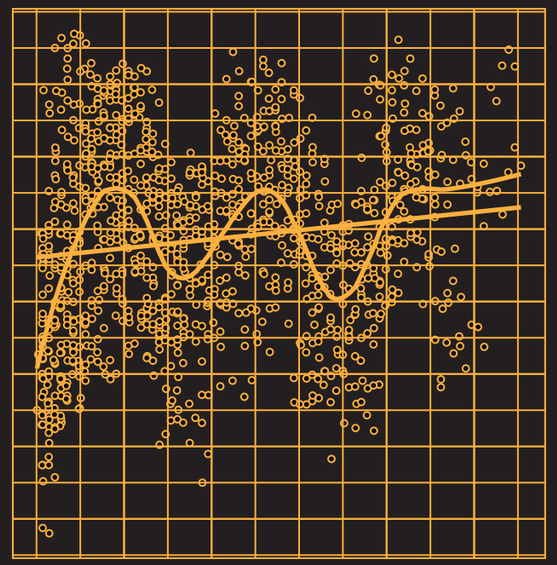
Microbial Macroecology

understanding microbial community pattems using phylogenetic and multivariate statistical tools

Albert Barberán Torrents



CSIC
Consejo Superior de Investigaciones Científicas

PhD Thesis 2012

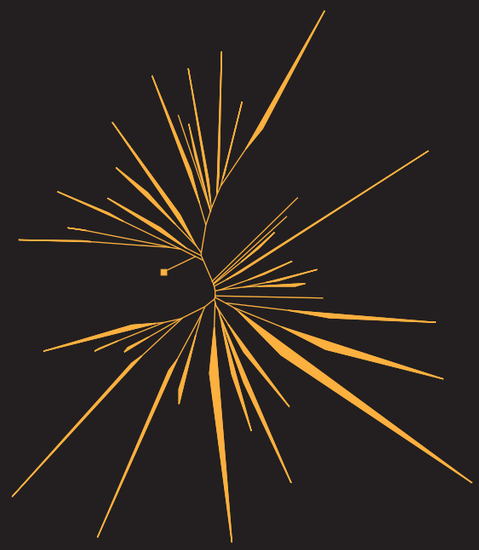
Microbial Macroecology

A. Barberán

Microbial Macroecology

understanding microbial community patterns
using phylogenetic and multivariate statistical tools

Albert Barberán Torrents



PhD Thesis
University of Barcelona

Universidad de Barcelona
Facultad de Biología

Microbial Macroecology

Macroecología Microbiana

Albert Barberán Torrents

Tesis Doctoral

Universidad de Barcelona

Facultad de Biología

Programa de Doctorado de Biodiversidad

Microbial Macroecology

understanding microbial community patterns using
phylogenetic and multivariate statistical tools

Memoria presentada por Albert Barberán Torrents
para optar al Título de Doctor en Biología

Albert Barberán Torrents

Centro de Estudios Avanzados de Blanes (CEAB)
Consejo Superior de Investigaciones Científicas (CSIC)

Barcelona, Junio de 2012

El director de la tesis
Dr. Emilio Ortega Casamayor
Inv. CEAB–CSIC

La tutora de la tesis
Dra. Marisol Felip Benach
Prof. UB

A Fausto i al meu ellvi,
por las lecciones y por las historias.

Agradecimientos

Ni el libro cerrado da sabiduría, ni el título por sí solo da maestría.

Refranero popular español

Dar las gracias es obscenamente fácil; tanto como pedir perdón. Por el contrario, hacer memoria y recordar las causas siempre ha requerido la ayuda de instancias superiores. Aunque durante el transcurso de esta tesis haya puesto a prueba mi acervo de ardides, la musa no me contará la historia de una tesis errabunda ni hablará de llanto ni crujir de dientes. Sin embargo, tantas mañas he usado en esta odisea que bien podría llamarse una tesis multidisciplinar.

Es de rigor comenzar agradeciendo el trabajo de mi jefe y director de tesis Emilio Casamayor. Me has conducido hasta aquí con ingenio y arte, a la vez que creabas un equipo de gran proyección en el que he disfrutado del trabajo y la convivencia con el resto de integrantes. Valoro el esfuerzo que me has dedicado y espero que estés orgulloso de los frutos cosechados de este mutualismo que establecimos hace ya más de cuatro años. Agradezco especialmente las discusiones sobre todo lo que rodea a la ciencia y que me hayas enseñado las reglas del juego.

Aunque nuestro grupo de investigación no ha funcionado exactamente como un todo orgánico, siempre hemos sido una *colla* muy bien avenida. Agradezco a JC los sudores en las campañas de los Pirineos y en las saunas, y a Antoni, el ser un verdadero camarada con el que he trabajado codo con codo y pantalla con pantalla. *Desitjo que aquesta tesi doctoral sigui digna de tu, perquè l'hi has dedicat molt d'esforç*. Hemos formado una tríada cohesionada por la ciencia y los efluvios: lo que ha unido la *chickipedia* que no lo separen los hombres (ni las mujeres).

Doy las gracias también a Xevi *pels ànims, l'ajut i l'exemple de les coses ben fetes*, y a Maria Vila-Costa por sus capciosas preguntas sobre el Mantel test. Tampoco debo olvidar mencionar a los miembros pretéritos: la Hierbas (que está embarazada), Pierre Galand y Laura Alonso. Aunque tenga espíritu de viejo cínico le deseo suerte a la savia nueva: a Tomàs, *et dono ànims perquè*

coronis la tesis tal com corones muntanyes, y a Claudia, gracias por escucharme, por abrirme los ojos, por llevarme siempre la contraria y por tus intentos de enseñar a bailar salsa a un catalán ñoño. Te deseo éxito en la tesis y os deseo mayor éxito en lo otro.

Gracias a las estancias de la beca FPU he podido ampliar mis miras por otros países. Agradezco la labor de acogida científica y personal de los grupos de investigación de Brendan Bohannan en la Universidad de Oregon, de Rampal Etienne en la Universidad de Groningen y de Noah Fierer en la Universidad de Colorado. La lista de personas con las que de una forma u otra me crucé es bastante extensa pero pecaría si no me acordase de Hélène Morlon, James O'Dwyer, Liz Perry, Kathryn Docherty, Fernando Encinas, Kelly Ramirez, Bob Bowers, Scott Bates o Chris Lauber. *The condor is coming back soon!*

Danke schön, Steffi. Sé que no existen palabras de gratitud suficientes que restañen la ingratitud de mis acciones. Desde dentro del CEAB, me guiaste la mirada fuera, lejos. Me presentaste a la escalada y al larguirucho de Hans. Tú me lo diste todo y yo fracasé. Siempre admiraré tu capacidad (como dijo tu compatriota) de luchar con monstruos y no convertirte en uno de ellos.

Da la impresión que agradecer al final del trayecto es una forma de despedida; por eso me es tan extraño darte las gracias, Carmen. No me despido porque sé que no te perderé nunca. Te seguiré acosando con mis tribulaciones, seguiré pidiéndote chorizo de tu madre, seguiré usufructuando tu televisor, seguiré deleitándome al hablar contigo y ver cómo baila tu inteligencia, y cuando mi inflado ego no me lo impida, te seguiré escuchando. Conoces buena parte de mis miserias y mis virtudes, de lo que enseño, de lo que escondo y de lo que escondo enseñándolo; en tus manos me encomiendo.

Si una tesis doctoral es el culmen de toda una vida de aprendizaje; es menester que también agradezca la labor de mis maestros de escuela y de mis profesores en la universidad. Lugar preeminente lo ocupa el Profesor Ricard Guerrero que me abrió las puertas de la investigación científica como placer y oficio. En su laboratorio, Laura Villanueva me hizo de jefa y Javier del Campo (del que estoy seguro que mandaría a la hoguera toda la obra de Prat de la Riba si así se salvase medio párrafo de Mishima), de Pistorius.

No sólo se aprende *de*, sino también *con*. Yo aprendí con mis amigos jinetes: Chema (Guerra), Borja (Muerte) y Manolo (Hambre). Juntos hicimos nuestras primeras lecturas serias, hablamos de política (antes de que se banalizase en los muros de Facebook), viajamos, derribamos árboles y vimos la esplendorosa Cibeles. En definitiva, prietamos y cabalgamos. En la Universidad también conocí a Maya, y con ella, me conocí más a mí mismo. Tanto ella como su familia me apoyaron sobremanera, especialmente durante el primer año de tesis mientras esperaba la maldita resolución de las becas.

Tras la Universidad recabé en el CEAB. Espero que en esta lista de agra-

decimientos ceabinos, aunque no estén todos los que son, sean todos los que están. A Fede Bartumeus, por su ejemplo de carrera científica y por prestarme sus archivos LaTeXpara la composición de esta tesis; a David Alonso, por su locura entusiasta y por sus acertados comentarios; a Miquel Ribot, porque es catalán y del Barça (como su compinche Peipoch); a la Riesgo, por sus charlas y porque cuentan los mentideros que fue una de las mejores estudiantes del CEAB; a Sergi Taboada por su guía en la elaboración e impresión de tesis; a Xavi Torras, porque sería el mejor de los suegros; a Magda, por su espíritu dicharachero; a Uri, por lo bien que nos quedan las faldas; a Tatiana, porque incluso la hiel me alimenta; a Clara y a Edu, porque vuestra iniciación en la escalada fue muy divertida, y a Gemma Agell y a Rocío, porque son unas grandes conversadoras. En general y a nivel científico, agradezco al CEAB el surtido de ejemplos de doctorandos y científicos a imitar, y también algún que otro contraejemplo a evitar.

Creo que algún ruso barbudo escribió aquello de que todas las familias felices se parecen, mientras que las desdichadas lo son cada una a su modo. El ruso sabía mucho, puede que demasiado, yo sólo sé que mi familia me habilitó una infancia cuya evocación me es muy agradable, y me dejó ser adulto cuando así lo necesité. El resultado de esta transición creo que no desagrada del todo a mis padres, o así me lo hacen sentir, tanto por su cariño como por su infatigable mecenazgo. Agradezco infinitamente la continua devoción de mis padres y mi abuela hacia su *panxacontent* hijo y nieto único, y si es por obligación de sangre, bienvenida sea y que no decaiga. Esta tesis es el resultado de mi curiosidad, la misma que me impulsó en el año 1994 a pedir como regalo el Libro Guinness de los Récords en sustitución de un viaje a EuroDisney. El desenlace de mis ansias mozas por aprender nombres de dinosaurios es esta tesis. Juzgad familia, si el fin justifica los medios y recordad que si alguna vez me descarrío; al diablo mismo, Dios fue quien lo hizo. Doy también gracias a mis primos, mayores y pequeños, por regalarme unas Navidades harto musicales donde mis dotes innatas para la zambomba se depuran año tras año.

De mi otra familia, la voluntaria, es un honor destacar los nombres de dos compadres de correrías que me acompañan desde la guardería, pasando por escuela e instituto. Godino, excelso hedonista, debo darte las gracias a ti por tu inquebrantable amistad, a tus padres y a tu buhardilla. Y Ramón, sé que lo sabes, pero no hay persona más lenta comiendo pizza que tú. Os prometo que no faltarán momentos para brindar por nuestros recuerdos y forjar de nuevos.

Me has acompañado en este corto tramo final y ya es mucho lo que debo agradecerte. Las ranas croan tu nombre y la orquídea no te olvida, ¿quién soy yo para no hacerles caso si me pirro por el pulpo a la gallega? Ojalá que todos mis defectos y mis zapatillas de Julio Iglesias no malbaraten la ilusión que me he propuesto insuflarte. Irene, prometo estarte agradecido.

Dos agradecimientos puntuales pero de gran relevancia en mi devenir: a *Muchacha Nui*, por su seriedad en torno a esta broma absurda llena de ruido y furia, y a la voz de Billie Holiday, porque en ocasiones consigue que olvide la dolorosa falsedad de sus letras.

No quisiera desaprovechar el beneplácito para pedir perdón (ya que es sencillo) a todos los que habéis sufrido mis histriónicas estridencias, mis gracias sin gracia, mi verborrea, mi inoportuna bocachancla, mis comentarios molestos, mi odiosa pedantería y mi egoísmo congénito. Y ya que se me permite (o eso creo), quiero señalar los obstáculos en el camino, que es tan de justicia como dar las gracias. Me ha estorbado (y me estorba) la burocracia; ese ente que olvida el porqué de su existencia, consume esfuerzos de forma cancerosa y nos transforma en esKarabajos.

El lugar de honor final en este catálogo de agradecimientos no podría ser ocupado por nadie más que por vosotros. Los que me habéis acompañado y me acompañaréis hasta mi muerte. Puede que a los únicos que he sabido amar. Gracias a vosotros no podré ser un caballo enjoyado, un asno con diadema o un membrillo con birrete doctoral. Vuestra llamada muerta me hace sentir vivo y os homenajeo diseminándoos por aquí y por allá para que algún avezado os reconozca. Por vosotros sigo buscándole nombre al fuego infinito.

Junio de 2012

Contents

1
Resumen

1.1 Introducción general

El estudio de los microorganismos en cultivo puro ha propiciado el desarrollo de la genética, la bioquímica y la biotecnología (Kluyver & Van Niel, 1956). Sin embargo, la ecología ha permanecido reticente a incorporar a los microorganismos en su acervo teórico y experimental, principalmente debido a las dificultades metodológicas para observar a los microbios en la naturaleza, y como resultado de los caminos divergentes que han trazado las disciplinas de la microbiología y la ecología general (Prosser et al., 2007). Mientras que los ecólogos han sido tradicionalmente muy adeptos a las aproximaciones teóricas y holísticas (e.g. Margalef, 1963), la investigación microbiológica se ha basado principalmente en la perspectiva reduccionista. No obstante, desde el advenimiento de las técnicas moleculares del gen ribosómico, los ecólogos microbianos son capaces de indagar la distribución de los microbios en el ambiente natural (Pace, 1997). Los datos derivan típicamente de la secuenciación parcial del gen de la subunidad pequeña del ribosoma que se utiliza como marcador taxonómico (Woese, 1987) con el fin de describir la composición, la diversidad y los patrones de variación de las comunidades microbianas.

En general, la distribución geográfica de los microorganismos se considera que se fundamenta exclusivamente en explicaciones puramente ecológicas. Esta suposición queda expresada mediante la máxima *todo está en todas partes; pero el ambiente selecciona* (Baas Becking, 1934). Este principio no implica la ausencia de patrones biogegráficos, sino que la geografía suele ser irrelevante debido al gran potencial dispersivo y a los enormes tamaños poblacionales de los microbios.

Una manera de abordar el problema en torno a la complejidad de las comunidades microbianas es adoptar aproximaciones holísticas e incrementar la escala de observación (Solé & Bascompte, 2006; Maurer, 1999). Ya que la impredecibilidad local es generalmente la característica más predecible de los sistemas ecológicos complejos, cuando se aumenta la escala de descripción, la variabilidad se reduce y la predicibilidad se incrementa (Levin, 1992). La disciplina de la macroecología busca ampliar el alcance de la ecología a mayores escalas espaciales y temporales mediante análisis estadísticos comparativos en lugar de métodos experimentales (Brown, 1995). Por lo tanto, se alcanza un mayor potencial de generalización y síntesis sacrificando la delineación detallada del fenómeno bajo estudio. Las comunidades microbianas son el sistema ideal para estudiar patrones macroecológicos porque el número total de células procariotas es astronómico (en torno a 10^{30}), y porque la diversidad microbiana se cree que podría llegar a ser del orden de 10^{6-9} (Whitman et al., 1998; Curtis et al., 2002).

Los ecólogos suelen distinguir tres perspectivas básicas para clasificar los principales factores que influencian los patrones de composición y diversidad de las comunidades biológicas. En primer lugar, la clásica explicación basada en nichos ecológicos otorga relevancia a los filtros ambientales locales y al principio de exclusión competitiva (Chase & Leibold, 2003). Por el contrario, la segunda perspectiva postula que la formación de las comunidades es un proceso fundamentalmente neutro basado en la dispersión y en procesos estocásticos (Hubbell, 2001). Finalmente, la tercera aproximación enfatiza el papel de los factores históricos, principalmente especiación y dispersión a escala regional, sobre los procesos locales (Ricklefs, 1987). Por lo tanto, la tarea primordial de la ecología de comunidades consiste en desentrañar la influencia relativa de los procesos de nicho, neutros e históricos. No se trata de una tarea sencilla ya que los patrones y los procesos subyacentes se interrelacionan a diferentes escalas, y ningún mecanismo explica las regularidades observadas a todas las escalas.

Es necesario la comprensión de los patrones de las comunidades microbianas para poder interpretar la organización funcional de los ecosistemas y para ser capaces de predecir la respuesta al cambio global (Fuhrman, 2009). Sin embargo, la investigación de patrones en comunidades microbianas data de fechas muy recientes. La cuestión de fondo que se plantea es si los patrones para microorganismos difieren radicalmente de las regularidades observadas en organismos macroscópicos (i.e. plantas y animales). Aunque esta cuestión está lejos de ser resuelta, la biología de los microbios presenta ciertas particularidades que requieren una atención especial. En lugar preeminente se encuentra el problema del concepto de especie en microorganismos asexuales. Las bacterias y las arqueas no permiten una clasificación taxonómica consis-

tente, ni si su tratamiento debe ser como grupos discretos o continuos, y además ponen en duda el patrón arbóreo de evolución biológica (Rosselló-Mora & Amann, 2001). La evolución microbiana ocurre a un ritmo mucho mayor que en macroorganismos, por lo que se ha propuesto que la gran diversidad de microbios se debe a una alta tasa de especiación unida a una baja tasa de extinción (Dykhuizen, 1998). Los dos mecanismos clave para comprender los patrones de distribución y diversidad microbianos son la dispersión y la capacidad de latencia. Ambos son mecanismos alternativos para reducir la variabilidad ambiental experimentada, y son fundamentales para la formación de la denominada "biosfera rara" (Pedrós-Alió, 2006).

Como se ha mencionado previamente, los sistemas ecológicos son el resultado de una combinación de procesos deterministas y estocásticos, junto con restricciones históricas. El más importante de dichos factores históricos es el hecho de que las especies no son entidades individuales, sino que sus semejanzas funcionales y ecológicas están conformadas a partir de patrones de ascendencia comunes (Felsenstein, 1985). En el campo de la ecología de comunidades, sólo recientemente los investigadores han incorporado dichas relaciones, representadas como filogenias, en sus análisis (Webb et al., 2002). Actualmente, las técnicas filogenéticas permiten determinar: (i) dónde se acumula la diversidad biológica (Faith, 1992) y cómo ésta está intrínsicamente estructurada (Webb, 2000; Helmus et al., 2007), y (ii) cómo la β-diversidad filogenética (i.e. la similitud entre comunidades basada en la historia evolutiva) está distribuida a lo largo de gradientes ambientales (Lozupone & Knight, 2005; Bryant et al., 2008). Como recalcaba constantemente el profesor Ramon Margalef, la ecología es el escenario donde la evolución desarrolla su función (Margalef, 1963). Como consecuencia, incorporar la topología filogenética a la ecología de comunidades es revelador porque permite dirigir las preguntas ecológicas hacia un contexto evolutivo.

1.2 Estructura, objetivos y resultados

Esta tesis trata de demostrar que los patrones ecológicos de comunidades microbianas son susceptibles de ser analizados mediante la combinación de técnicas filogenéticas y herramientas de estadística multivariante. El uso de técnicas filogenéticas permite solventar, o al menos paliar, el hecho de la no independencia de los organismos vivos debido a la ascendencia común (Felsenstein, 1985). Con la información ambiental adicional (como reflejo del determinismo abiótico) y la información espacial (como amalgama de eventos históricos y de dispersión), es posible explorar los posibles mecanismos que subyacen a la estructura y a la diversidad de las comunidades microbianas (Legendre & Legendre, 1998).

1.2.1 **Primera parte:** Meta-análisis filogenético de secuencias microbianas del gen 16S rRNA a escala global

Actualmente, la composición y la diversidad de las comunidades microbianas pueden ser evaluadas mediante aproximaciones moleculares; incluso cuando se desconocen los atributos funcionales y ecológicos de la mayor parte de sus miembros. Las cuatro capítulos incluidos bajo esta sección utilizan métodos comparativos a escala global: los dos primeros con comunidades acuáticas bacterianas y los dos siguientes analizando comunidades de arqueas.

Capítulo 4: Patrones de β-diversidad y estructura filogenética de comunidades en metacomunidades de bacterioplancton a escala global [1]

Este trabajo intenta identificar patrones filogenéticos de comunidades de grupos de bacterias abundantes en el plancton (*Alpha-*, *Beta-*, y *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, y *Bacteroidetes*) en sistemas acuáticos superficiales distribuidos globalmente (34 localizaciones -tanto lagos como mares- que suman 4500 secuencias del gen 16S rRNA). En cada localización se estimó la riqueza de grupos bacterianos y sus relaciones genéticas mediante una aproximación de filogenia de comunidades con el fin de (i) explorar aquellos procesos ecológicos compatibles con los patrones filogenéticos observados, y (ii) desentrañar los efectos espaciales y del ambiente en los patrones de β-diversidad para los diferentes grupos de bacterias.

Las masas acuáticas terrestres presentaron significativamente una mayor riqueza y diversidad de grupos bacterianos que las localizaciones marinas. Éstas, por su parte, presentaron un porcentaje mayor de emplazamientos agregados filogenéticamente. Los análisis de β-diversidad revelaron pautas con-

[1] Ver publicación completa en Barberán & Casamayor (2010).

trastadas en función tanto de la composición de sales (aguas marinas/masas acuáticas terrestres) como de la concentración salina. Se observó que dichos patrones para el grupo *Bacteroidetes* estaban estructurados según la concentración salina, mientras que en *Alphaproteobacteria* y *Gammaproteobacteria* estaban controlados según la composición de sales. Los grupos de *Actinobacteria*, *Betaproteobacteria* y *Sphingobacteria* fueron raramente detectados en aguas marinas y aguas continentales salinas. En general y pese a la falta de una profusión de datos contextuales, la similitud ambiental fue más determinante que la espacial para la distribución de la β-diversidad de bacterioplancton. No obstante, se detectó una señal geográfica para algunos grupos (i.e. *Actinobacteria*, *Beta-*, y *Gammaproteobacteria*) en aguas continentales. A grandes rasgos, los análisis indicaron diferencias entre grupos filogenéticos y reflejaron patrones interesantes para investigaciones subsiguientes en torno a la formación de comunidades microbianas.

Capítulo 5: El hipolimnion euxínico en aguas dulces continentales promueve el endemismo bacteriano[2]

Los dominios *Bacteria* y *Archaea* representan la vasta mayoría de la biodiversidad de la Tierra. No obstante, la interacción entre los procesos ecológicos y evolutivos en el mundo microbiano permanece desconocida. En este estudio, se han explorado los patrones de comunidades planctónicas de bacterias que habitan lagos estratificados con capas óxicas/anóxicas y euxinia. Se examinó si esta estratificación vertical es promotora del endemismo en capas profundas mediante el análisis de secuencias del gen ribosómico del 16S.

La similitud en la composición bacteriana mostró que las comunidades de la misma capa de agua eran más parecidas entre lagos que las comunidades de diferentes capas del mismo lago. Además, el hipolimnion anóxico presentó mayor β-diversidad que el epilimnion óxico. Una mayor β-diversidad puede ser atribuible a una baja dispersión y escasa conectividad entre territorios. Paralelamente, mientras que las aguas superficiales exhibieron un componente espacial significativo, en el caso de las capas profundas, el componente significativo fue el ambiental. Por lo tanto, diferentes mecanismos ecológicos actúan simultáneamente en la misma masa de agua. En general, el endemismo en bacterias es probablemente más común que lo supuesto con anterioridad, particularmente en hábitats de aguas dulces aislados y ambientalmente heterogéneos.

[2]Ver publicación completa en Barberán & Casamayor (2011).

Capítulo 6: Patrones ecológicos globales en *Archaea* sin cultivar[3]

Al aplicar una aproximación filogenética global a organismos *Archaea* no cultivados se han revelado patrones de comunidades definidos a lo largo de amplios gradientes y tipologías ambientales. El análisis se ha basado en unas 2000 secuencias del gen 16S ribosómico de arqueas provenientes de 67 localizaciones. Dichas secuencias fueron agregadas al 97 % de identidad, clasificadas en siete tipos de hábitats, y analizadas tanto con UniFrac (para explorar la historia filogenética compartida) como mediante árboles de regresión multivariantes (que consideran la abundancia relativa de los diferentes linajes). Ambas perspectivas apuntaron a la salinidad como el factor regulador principal a escala global. Las chimeneas hidrotermales y los hábitats planctónicos continentales se postularon como los mayores reservorios de diversidad de arqueas y, por lo tanto, como ambientes prometedores para el descubrimiento de nuevos linajes. Por el contrario, los suelos exhibieron una mayor agregación filogenética, resultado de la presencia de filotipos relacionados estrechamente. Se detectaron diferentes linajes indicadores para los distintos hábitats, algunos de los cuales desconocidos a nivel ecológico. Según los análisis de distribución de comunidades, las chimeneas hidrotermales parecen ser uno de los primeros hábitats colonizados por arqueas. En resumen, este estudio proveyó de soporte ecológico a la nomenclatura arbitraria de *Archaea*, a la par que desveló aspectos filogeográficos relevantes en su biología.

Capítulo 7: Ecología filogenética de clados sin cultivar del reino *Euryarchaeota*[4]

A pesar de su extensa distribución y su gran diversidad filogenética, los microorganismos continúan siendo unos grandes desconocidos a nivel ecológico. Con el fin de ahondar en la distribución ambiental y la historia evolutiva se aplicó una aproximación de ecología filogenética basada en las secuencias del gen 16S rRNA a dos grupos del reino *Euryarchaeota*, Lake Dagow Sediment (LDS) y Rice Cluster-V (RC-V). La historia evolutiva inferida indicó que ambos grupos han sufrido evolución específica en cada ambiente, con algunos notables eventos de transición entre hábitats. Comparado con otros grupos microbianos de arqueas, ambos grupos presentaron remarcables niveles de diversidad genética posiblemente fomentada por su adaptabilidad ambiental y la heterogeneidad de las masas de agua continentales donde medran. Los diversificación a lo largo de la historia filogenética se concentró tanto en los instantes iniciales como en los más recientes. Para la mayoría de microorga-

[3] Ver publicación completa en Auguet et al. (2010).
[4] Ver publicación completa en Barberán et al. (2011).

nismos, la diferenciación genética y fisiológica que habilita la explotación de los recursos ambientales sigue desconocida. La inferencia de la historia evolutiva a partir de filogenias moleculares basadas en el gen ribosómico del 16S permite una perspectiva ecológica con la que escudriñar las estrechamente entrelazadas relaciones entre los linajes, el ambiente y la contingencia histórica en el mundo microbiano.

1.2.2 **Segunda parte:** Ecología de comunidades bacterianas en lagos pirenaicos y la influencia de las deposiciones de polvo sahariano

Una vez asegurada la viabilidad de la aproximación filogenética usada en la **Primera Parte** y de los análisis de redes desarrollados en el **Capítulo 10** de esta tesis, el foco de atención se desplazó de los análisis comparativos a escala global a el análisis de datos originales a escala regional. El ecosistema escogido fue el formado por los lagos de alta montaña localizados en el área de los Pirineos catalanes en torno al Parque Nacional de Aigüestortes y el lago de San Mauricio. El estudio de este sistema se remonta a inicios de los años ochenta con la labor pionera del profesor Ramon Margalef (Catalan et al., 2006). Este cambio de escala espacial facilitó la investigación del impacto de procesos globales como la deposición de polvo atmosférico entre continentes en las comunidades microbianas a escala local y regional. El segundo de los trabajos incluídos en esta sección es fruto de una colaboración con colegas de la Universidad de Colorado.

Capítulo 8: Una perspectiva filogenética de las comunidades bacterianas en lagos pirenaicos de alta montaña [5]

Las regiones montañosas suelen tener un variado mosaico de pequeñas masas de agua, y por lo tanto, conforman un adiente modelo biogeográfico de ecosistemas cercanos geográficamente pero de gran heterogeneidad ambiental. Además, el carácter aislado y prístino de estos lagos los hace excelentes centinelas y testigos de los cambios climáticos. En este estudio se muestrearon dieciocho lagos pirenaicos basándose en un criterio de selección que maximizase la variación ambiental, y se secuenció el gen ribosómico de la subunidad 16S bacteriana con el objetivo de describir la composición filogenética, su distribución espacial y los patrones de β-diversidad.

Los resultados obtenidos demuestran que las comunidades bacterianas de los lagos de los Pirineos presentan una mayor abundancia de *Betaproteobacteria*

[5] Manuscrito en preparación.

y una menor abundancia de *Alphaproteobacteria* que otros ecosistemas planctónicos de agua dulce. Se observó que los patrones de β-diversidad estaban fundamental estructurados por el ambiente, siendo el gradiente de pH el más determinante. De manera interesante, se detectó una relación positiva entre el área de los lagos y la diversidad filogenética, con una pendiente acorde a organismos planctónicos de alta capacidad dispersiva. La aproximación filogenética usada en este trabajo, por la cual se incorporan los patrones de ascendencia común en los análisis ecológicos de comunidades, resultó ser una herramienta idónea para los estudios de conservación que recurrentemente se han descuidado de los microorganismos.

Capítulo 9: Estructura y patrones temporales de las comunidades microbianas transportadas entre continentes vía polvo atmosférico proveniente del Sahara [6]

El polvo originado en los grandes desiertos puede viajar largas distancias y ser dispersado sobre superfícies de cientos de miles de kilómetros cuadrados. El proceso de la deposición de polvo está viéndose incrementado a consecuencia del cambio global y constituye una fuente de nutrientes y también un mecanismo para la dispersión microbiana intercontinental. Los lagos alpinos oligotróficos son perfectos centinelas de la magnitud de este proceso, del destino de los microbios colonizadores, y de la calidad microbiológica de la atmósfera global.

En este trabajo, se estudió la composición de las comunidades microbianas de tres hábitats potencialmente conectados atmosféricamente: el origen de las plumas de polvo (i.e. arena del desierto del Sáhara), el transportador atmosférico (i.e. la deposición que llega a la región montañosa de los Pirineos Centrales durante un periodo de tres años) y el colector natural de los microorganismos aerotransportados (i.e. la interfase aire-agua de los lagos de alta montaña). Se ha descrito la composición taxonómica, los patrones de β-diversidad y la coincidencia en OTUs entre los tres hábitats. El porcentage de OTUs compartidos fue del 10 % del número total de OTUs. Las comunidades microbianas de la deposición atmosférica presentaron un claro patrón temporal. En resumen, el estudio sugiere que existe una variabilidad consistente y predecible en la dinámica y la distribución de las comunidades microbianas aerotransportadas, y que las formas celulares viables con el potencial de colonizar nuevos ambientes no sólo se restringen a las formas esporuladas.

[6] Manuscrito en preparación.

1.2.3 **Tercera parte:** Nuevas aproximaciones ecológicas a datos masivos de metagenómica y de pirosecuenciación

Recientemente, el estudio de la ecología de comunidades microbianas ha avanzado considerablemente debido al advenimiento de nuevas técnicas moleculares como la pirosecuenciación (Sogin et al., 2006) y la metagenómica (Venter et al., 2004) que generan multitud de datos genéticos. No obstante, el ritmo de acumulación de información sobrepasa el ritmo al cual los científicos pueden interpretar dichos datos. Bajo esta sección se agrupan dos trabajos en los que se toman prestadas técnicas de la ciencia de sistemas complejos y de la ecología general con el fin de adjudicar sentido ecológico a grandes conjuntos de datos de información genética. El primero de los estudios es fruto de una cooperación con colegas de la Universidad de Oregón, mientras que el segundo surgió tras una estancia de colaboración en la Universidad de Colorado.

Capítulo 10: Análisis de redes para la exploración de patrones de co-ocurrencia en comunidades microbianas de suelos [7]

La exploración de extensos conjuntos de datos generados mediante técnicas de secuenciación masiva requiere nuevas aproximaciones analíticas para escudriñar más allá de los inventorios descriptivos en torno a la composición y la diversidad de las comunidades microbianas. Con el objetivo de investigar potenciales interacciones entre clados de microorganismos, el análisis de redes basados en patrones de co-ocurrencia puede ser una herramienta útil para descifrar la compleja estructura de estas comunidades.

En este trabajo, se aplicó el análisis de redes a un conjunto de datos del gen ribosómico del 16S (más de 160000 secuencias) obtenido por pirosecuenciación de 151 muestras de suelos. Se describió la topología de la red generada y se definieron clados generalistas y especialistas en función de su abundancia y presencia. Gracias a esta nueva perspectiva se pudieron apreciar patrones de co-ocurrencia no aleatorios y estrategias comunes a amplios niveles taxonómicos. En conclusión, se demostró el potencial de la exploración de correlaciones entre clados para la aprehensión de ciertas reglas ecológicas que guían el ensamblaje de las comunidades microbianas.

[7] Ver publicación completa en Barberán et al. (2012a).

Capítulo 11: Caracteres a nivel de comunidad como marcadores ecológigos en metagenomas microbianos [8]

El ritmo de recopilación de información generada por la técnica de la metagenómica está desacoplado de su interpretación ecológica elocuente. Nuevas aproximaciones analíticas basadas en la ecología de caracteres funcionales podrían ser de gran ayuda para solventar este desacoplamiento y extender dicha aproximación al nivel de comunidad en complejos conjuntos de datos genómicos. El objetivo de este estudio fue la exploración de un grupo de caracteres comunitarios que cubrían propiedades tanto nucleotídicas como genómicas en 53 muestras metagenómicas acuáticas provenientes de la expedición GOS.

Como resultado, se encontraron diferencias significativas entre el perfil de β-diversidad derivado del marcador taxonómico del gen ribosómico del 16S y el perfil funcional. Los caracteres analizados discriminaron entre ecosistemas marinos y entre océanos, por lo que se postulan como potenciales marcadores ecológicos. Además, algunas relaciones entre caracteres podrían ser usadas como señales particulares de hábitats o incluso como indicadores de artefactos durante el procesamiento de muestras. Como conclusión, la perspectiva analítica presentada puede ser fructífera para la interpretación de datos metagenómicos dentro de un riguroso marco ecológico.

[8] Ver publicación completa en Barberán et al. (2012b).

1.3 Conclusiones generales

A pesar de algunos acercamientos durante el siglo XX, los estudios ecológicos de microorganismos no forman históricamente parte de la ecología general, sino que son una subdisciplina de la microbiología (O'Malley & Dupré, 2007). En referencia a los patrones macroecológicos, existe la incertidumbre en torno a la unidad ecológica de todas las formas de vida o si, por el contrario, los microbios difieren radicalmente de animales y plantas. Aunque la ecología microbiana ha permanecido hasta ahora en un fase descriptiva, semejante a la ecología de plantas y animales durante el siglo XIX, es el momento de realizar contribuciones relevantes desde el campo de la ecología de microorganismos a la ecología general (Fierer, 2008). Muchos de los patrones que apoyan la evidencia de la distribución ambiental y geográfica no aleatoria de los microbios fueron estudiados durante décadas en comunidades vegetales y animales pero han sido ignorados por la microbiología hasta el advenimiento de las técnicas moleculares que permiten analizar la composición de las comunidades microbianas.

Durante la pasada década, la ecología microbiana lleva generando abundantes datos moleculares. Actualmente, la combinación de herramientas bioinformáticas y estadísticas junto con conceptos teóricos permite integrar a los microorganismos dentro del campo de la ecología general (**Capítulos 4, 5, 6, 7 y 10**). En paralelo a los estudios ambientales del gen de la subunidad pequeña del ribosoma, la nueva técnica de la metagenómica desafía a la comunidad científica con un gran volumen de datos en la intersección de las disciplinas de la microbiología, la genética, la ecología y la bioinformática. Por lo tanto, la metagenómica está descubriendo una nueva faceta de la complejidad de las comunidades microbianas al desplazar el foco de interés de la composición taxonómica a la composición de genes funcionales (**Capítulo 11**).

Los observatorios microbianos en ambientes prístinos y remotos se postulan como útiles centinelas del cambio global, y como testigos del transporte atmosférico a larga distancia. Un primer paso prospectivo en torno a los mecanismos de dispersión microbianos es el estudio de la deposición que viaja a altas altitudes y llega a los lagos de alta montaña (**Capítulos 8 y 9**).

Si los estudios comparativos a gran escala entre ambientes contrastados pueden ser una primera aproximación para comprobar los efectos de los límites a la dispersión en la estructura de las comunidades microbianas (**Capítulos 4 y 5**); estudiar la composición microbiana de la deposición atmosférica puede ser importante para entender el papel potencial de la dispersión a larga distancia y las estrategias de supervivencia de los microorganismos (**Capítulo 9**).

Aunque acumular información descriptiva sobre la composición microbia-

na ambiental es todavía necesaria (como se ha llevado a cabo en los **Capítulos 8 y 9**), todavía parece más necesario desarrollar técnicas para poder analizar los datos generados. Con este fin, se han tomado prestados métodos y conceptos de la ecología (**Capítulos 6 y 11**), de la biología evolutiva (**Capítulo 7**) y de la ciencia de los sistemas complejos (**Capítulo 10**) al campo de la ecología microbiana. Además, se ha propuesto un método con el potencial de describir, analizar y detectar posibles anomalías en datos metagenómicos **(Capítulo 11**).

Como culmen se propone un modelo neutro alternativo basado en las idiosincracias de las comunidades microbianas que puede ser el pilar y el marco para futuras investigaciones.

1.3.1 Dos modelos neutros para una teoría unificada de la biodiversidad

Entender la estructura jerárquica y compleja de la biodiversidad ("el barroco de la naturaleza", como solía referirlo Margalef) es una de las tareas más desafiantes de la ciencia moderna (Solé & Bascompte, 2006). Mientras que la teoría de nichos plantea que cada especie posee un conjunto de caracteres únicos que le permiten adaptarse a un ambiente abiótico y biótico concreto (Chase & Leibold, 2003); la teoría neutra centra su atención en procesos usualmente desdeñados como la deriva ecológica y la dispersión (Hubbell, 2001). En la formulación de Hubbell, todos los individuos de diferentes especies en la comunidad son estrictamente equivalentes en referencia a sus posibilidades de reproducción y muerte. Por lo tanto, las evidentes y archiconocidas diferencias entre especies son irrelevantes para la predicción de patrones a gran escala, siendo el punto crucial el determinar el alcance real de las diferencias funcionales. Las comunidades ecológicas pueden parecer neutras debido a su complejidad, con patrones que emergen de un proceso estadístico de intrincada casuística. Por consiguiente, la teoría neutra se asemeja a la teoría cinética de los gases: es una teoría ideal (Alonso et al., 2006).

El modelo descrito por la formulación original de Hubbell provoca ciertamente insatisfacción a cualquier ecólogo microbiano ya que no contempla las idiosincracias propias de los microorganismos. La teoría neutra necesita de dos modelos (uno para macroorganismos y otro para microorganismos) para dar cuenta de toda la diversidad biológica. Aunque en ambos modelos los mismos mecanismos evolutivos y de dispersión dan forma a la composición, la diversidad y la dinámica de las comunidades naturales; existe una transición primordial en las escalas en las cuales dichos mecanismos operan.

En el modelo modificado para dar cabida a las comunidades microbianas, la escala regional se desestima debido al gran potencial dispersivo de los microbios, y la escala global gana en relevancia. Si en el modelo neutro de

Hubbell para macroorganismos, las escalas local y regional están acopladas a través de migración unidireccional (i.e. colonización desde la metacomunidad hacia la comunidad local); en el modelo para microbios, a los individuos les es permitido dispersarse a largas distancias y, por tanto, formar parte del acervo global (la llamada "biosfera rara"; Pedrós-Alió, 2006) como resultado de sus capacidades dispersivas y de latencia (Locey, 2010). De este modo, los organismos pueden abandonar las comunidades locales por extinción fortuita, o por dispersión a larga distancia para formar parte del "banco de semillas" global. Con el fin de parametrizar la elevada tasa de especiación de los microorganismos en comparación con la de plantas y animales (Dykhuizen, 1998), la especiación tiene lugar a escala local. Aunque para la descripción de ambos modelos se ha utilizado una formulación en términos discretos, el problemático concepto de especies para microbios asexuales sugeriría tratar la especiación como un proceso continuo y no discretizado (Rosselló-Mora & Amann, 2001).

En resumidas cuentas, se ha intentado sugerir que la ecología de comunidades microbianas debería aventajar la socorrida máxima de Baas Becking, *todo está en todas partes; pero el ambiente selecciona*, hacia una teoría estocástica simple formulada sobre mecanismos ecológicos y evolutivos conocidos.

1.3.2 Conclusiones

Las conclusiones generales de esta tesis son las siguientes:

- Tres procesos principales explican los patrones de comunidades: deterministas, estocásticos e históricos. Es esencial la incorporación de la información filogenética en los análisis de comunidades para tener en cuenta la estructura histórica de los organismos vivos (**Introducción**).
- Ciertas características idiosincráticas de los microorganismos (i.e. alta especiación, alto potencial dispersivo y la capacidad de latencia) son fundamentales para entender los patrones de comunidades (**Introducción**).
- Las masas de agua aisladas y con elevada heterogeneidad ambiental presentan mayor diversidad microbiana (**Capítulos 4 y 5**).
- A escala global y con grupos taxonómicos amplios, el filtro por el ambiente es más relevante que los procesos geográficos para explicar la estructura de las comunidades microbianas (**Capítulos 4, 5 y 6**).
- Las comunidades de arqueas presentan patrones definidos a escala global (**Capítulo 6**).

- El estudio de clados de microorganismos sin cultivar se ve mejorado al incorporar su patrón temporal de diversificación (**Capítulo 7**).
- Los ecosistemas acuáticos continentales son ambientes con gran potencial para el descubrimiento de nueva diversidad microbiana (**Capítulos 4, 5, 6 y 7**). En concreto, la heterogeneidad ambiental en lagos pirenaicos promueve una elevada diversidad filogenética (**Capítulo 8**).
- La diversidad filogenética de las comunidades bacterianas de lagos pirenaicos se adhiere a patrones biogeográficos conocidos (**Capítulo 8**).
- La deposición de polvo atmosférico actúa de puente entre las escalas global y regional para las comunidades microbianas y su composición presenta un patrón temporal regular (**Capítulo 9**).
- El análisis de redes permite la exploración inicial de patrones de coocurrencia (**Capítulo 10**) y el análisis de caracteres funcionales, la descripción y la interpretación de datos metagenómicos (**Capítulo 11**).
- Los componentes taxonómico y funcional muestran diferentes instantáneas de las comunidades microbianas. Su comparación puede resolver la cuestión sobre la capacidad de adaptación (ya sea por variación funcional o taxonómica) de las comunidades microbianas (**Capítulo 11**).
- Un modelo neutro que otorgue preponderancia a la escala global debido a la dispersión a larga distancia y que sitúe la especiación a escala local parece más adecuado para comunidades de microorganismos que los formulados con anterioridad (**Conclusiones**).

Informe del director

El Dr. **Emilio Ortega Casamayor**, Investigador Científico del Centro de Estudios Avanzados de Blanes-CSIC, en calidad de Director de la tesis doctoral *Microbial Macroecology: Understanding microbial community patterns using phylogenetic and multivariate statistical tools* presentada por Albert Barberán Torrents para optar al título de Doctor dentro del programa de doctorado en Biodiversidad de la Universidad de Barcelona, hace constar que la participacion del aspirante a doctor en cada uno de los artículos presentados en esta memoria es la que se detalla en los párrafos siguientes. Así mismo constata que en siete de los ocho artículos científicos generados con esta tesis doctoral, el candidato es primer firmante de los trabajos y en ningún caso el resto de coautores ha utilizado, implícita o explícitamente, datos o resultados de estos trabajos para la elaboración de ninguna otra tesis doctoral. Seis de los ocho artículos se encuentran ya publicados y lo han sido en revistas internacionales de referencia de la especialidad pertenecientes al primer cuartil (Q1) SCI, tanto de la categoria ISI *Ecology* como de la categoría *Marine & Freshwater Biology*.

- **Artículo I**
  Barberán A, EO Casamayor (2010) Global phylogenetic community structure and beta-diversity patterns of surface bacterioplankton metacommunities. *Aquatic Microbial Ecology* 59: 1-10.

  Indicadores bibliométricos. IF (2010): 2.089; Cuartil: Q1; Categoría: MARINE & FRESHWATER BIOLOGY (posición 19 de 93); Citas SCI acumuladas: 22.

  En este trabajo se ha realizado una minería de datos de genes ribosómicos de bacterias lacustres y marinas presentes en las bases de datos y un análisis de datos mediante filogenia de comunidades y estadística multivariante. El candidato a doctor lideró desde el primer momento el diseño experimental, la aplicación de metodologías y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo y actuando como *corresponding author* del mismo. El artículo fue seleccionado como *Feature Article* ocupando un lugar prominente en el volumen de la revista.

- **Artículo II**
  Barberán A, EO Casamayor (2011) Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microbial Ecology* 61: 465-472.

  Indicadores bibliométricos. IF (2010): 2.875; Cuartil: Q1; Categoría: MARINE & FRESHWATER BIOLOGY (posición 12 de 93); Citas SCI acumuladas: 4.

  El candidato a doctor participó en el diseño experimental, y llevó a cabo la obtención de los datos, la aplicación de metodologías de análisis filogenético y estadístico, y la redacción del primer borrador del manuscrito.

- **Artículo III**
  Auguet JC, A Barberán, EO Casamayor (2010) Global ecological patterns in uncultured Archaea. *The ISME Journal* 4: 182-190.

  Indicadores bibliométricos. IF (2010): 6.153 ; Cuartil: Q1; Categoría: ECOLOGY (posición 7 de 130); Citas SCI acumuladas: 41. Highly cited paper (ISI WoK Essential Science Indicators).

  El candidato a doctor participó en el diseño experimental del trabajo, en la aplicación de las metodologías de análisis filogenético y estadístico desarrolladas en Barberán & Casamayor (2010) y participó activamente en la discusión e interpretación de los resultados. Este trabajo ha sido publicado en la revista de mayor índice de impacto de la especialidad de ecología microbiana.

- **Artículo IV**
  Barberán A, Fernández-Guerra A, Auguet JC, Galand PE, EO Casamayor (2011) Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota. *Molecular Ecology* 20: 1988-1996.

  Indicadores bibliométricos. IF (2010): 6.457; Cuartil: Q1; Categoría: ECOLOGY (posición 5 de 130); Citas SCI acumuladas: 2.

  El candidato a doctor participó en el diseño experimental, y llevó a cabo el tratamiento de los datos, la aplicación de metodologías de análisis filogenético y estadístico y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo y actuando como *corresponding author* del mismo.

- **Artículo V**
  Barberán A, EO Casamayor (manuscrito en preparación) A phylogenetic perspective on bacterial communities from high-altitude Pyrenean lakes.

Este trabajo es pionero en la descripción de la diversidad y estructura filogenética de las bacterias presentes en diferentes ambientes del sistema lacustre de los Pirineos centrales, y en los factores ambientales que modulan la estructura de las comunidades. Así mismo se aplica el concepto de riqueza filogenética de la comunidad a las predicciones de la teoría biogeográfica de islas entendiendo los lagos como islas rodeadas por un océano de tierra. El candidato generó los datos moleculares, lideró el tratamiento de los datos con la aplicación de metodologías de análisis filogenético y estadístico, y participó activamente en el desarrollo conceptual del trabajo y la redacción del primer borrador del manuscrito.

- **Artículo VI**
  Barberán A, J Henley, N Fierer, EO Casamayor (manuscrito en preparación) Structure and temporal patterns in microbial communities transported across continents during Saharan dust events.

  Este trabajo se ha realizado como una colaboración a raíz de una estancia predoctoral realizada por el candidato en el laboratorio del Dr. Noah Fierer en el Departamento de Ecología y Biología Evolutiva de la Universidad de Colorado (USA). Se aborda el incipiente campo de la ecología microbiana de la atmósfera y los mecanismos globales de dispersión de microorganismos. El candidato lideró el tratamiento de los datos, la aplicación de metodologías de análisis filogenético y estadístico a datos generados por secuenciación masiva de genes ribosómicos, participando activamente en el desarrollo conceptual del trabajo y la redacción del primer borrador del manuscrito.

- **Artículo VII**
  Barberán A, ST Bates, EO Casamayor, N Fierer (2012) Using network analysis to explore co-ocurrence patterns in soil microbial communities. *The ISME Journal* 6: 343-351.

  Indicadores bibliométricos. IF (2010): 6.153 ; Cuartil: Q1; Categoría: ECOLOGY (posición 7 de 130); Citas SCI acumuladas: 0.

  Este trabajo es fruto de una estancia predoctoral de 3 meses en el laboratorio del Dr. Noah Fierer en el Departamento de Ecología y Biología Evolutiva de la Universidad de Colorado (USA). El candidato participó activamente en el diseño experimental, y llevó a cabo el tratamiento de los datos, la aplicación de metodologías de análisis filogenético y estadístico por análisis de redes de datos obtenidos por secuenciación masiva de genes ribosómicos y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo y actuando como *corresponding*

author del mismo. Este trabajo ha sido publicado en la revista de mayor índice de impacto de la especialidad de ecología microbiana.

- **Artículo VIII**
  Barberán A, A Fernández-Guerra, B Bohannan, EO Casamayor (2012) Exploration of community traits as ecological markers in microbial metagenomes. *Molecular Ecology* 21: 1909-1917.

  Indicadores bibliométricos. IF (2010): 6.457; Cuartil: Q1; Categoría: ECOLOGY (posición 5 de 130); Citas SCI acumuladas: 1.

  Este trabajo es fruto de una estancia predoctoral de 3 meses en el laboratorio del Dr. Brendan Bohannan en el Centro de Ecología y Biología Evolutiva de la Universidad de Oregon (USA). El candidato participó activamente en el diseño conceptual del trabajo, y llevó a cabo el tratamiento de los datos, la aplicación de metodologías de análisis taxonómico y funcional de datos metagenómicos, discusión activa de los resultados y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo y actuando como *corresponding author* del mismo. La aplicación de conceptos de ecología de comunidades a datos metagenómicos es una idea pionera y el artículo fue seleccionado por el comité editorial de *Molecular Ecology* para aparecer en un número especial dedicado a DNA ambiental recientemente reseñado en la revista *Nature*.

El director de la tesis
Dr. Emilio Ortega Casamayor
Inv. CEAB-CSIC

Microbial Macroecology

2
General introduction

The study of microorganisms in pure laboratory culture has delivered fruitful insights into genetics, biochemistry and biotechnology (Kluyver & Van Niel, 1956). However, ecology has remained reluctant to incorporate microorganisms in its experimental and theoretical underpinnings (but see Jessup et al., 2004) mainly due to methodological difficulties in observing microorganisms in nature [1], and as a result of the different paths followed by the disciplines of microbiology and general ecology (Prosser et al., 2007). While ecologists have traditionally been very prone to theoretical and holistic perspectives (e.g. Margalef, 1963; Odum, 1969), microbiological research has relied on a reductionist approach. Understanding highly diverse microbial communities embedded in a complex environmental milieu with ecological and evolutionary processes operating at multiple spatial and temporal scales, certainly requires the borrowing of ecological concepts and theoretical tools because:

> No science has succeeded in understanding the structure and dynamics of a complex system from a reductionist approach alone. Physicists have not solved the three-body problem, molecular biologists have not recreated an organism from its chemical constituents in a test tube, and neurobiologists have not been able to explain memory and cognition in terms of interactions among neurons. (Brown, 1995, p. 15)

Notwithstanding, since the advent of 16S rRNA gene molecular techniques, microbial ecologists are able to question the distribution of microor-

[1] Knowledge about the diversity of microorganisms is minimal comparing to plants and animals, where perhaps 90-99% of all species are known. In contrast, it is estimated that less than 1% of microbial diversity is known. Whereas 11,000 species of *Bacteria* and *Archaea* have been formally described (http://www.bacterio.cict.fr); it is expected that the number of different bacterial species present on Earth ranges between 10^6 to 10^9 (Dykhuizen, 1998; Curtis et al., 2002).

ganisms in the environment (Pace, 1997). Data are tipically derived from sequencing a portion of the small-subunit rRNA gene, which is seen as a proxy for natural taxonomical units (Woese, 1987) in order to describe the composition and diversity of microbial communities, and how communities change across space, time, or experimental treatments. In the most recent years the study of complex microbial assemblages has advanced considerably due, in part, to methodological advances such as metagenomics and high-throughput DNA sequencing technologies that yield detailed information on the phylogenetic and functional composition of microbial communities (Venter et al., 2004; Sogin et al., 2006). Unfortunately, the rate of information collection by molecular techniques is by far outpacing the rate at which researchers can properly analyze and interpret the data.

Paraphrasing MacArthur & Wilson (1967, p. 181): "Microbial ecology has been blindly triggered by technical developments and hence, has long remained in a natural history phase, accumulating information about the biotic composition. Without doubt this descriptive phase will continue to be of fundamental importance, but the field of microbial ecology is presently entering a fascinating analytical and holistic phase."

2.1 A holistic approach to microbial community ecology

A holistic view of microbial ecology first needs to ascertain the geographic distribution of diversity (i.e. biogeography) before examining complex patterns and the underlying mechanistic processes. The birth of biogeography is usually associated with the rejection by Carl Linnaeus (1707-1778) of the biblical accounts of Noah's Ark, and the recognition by the Comte de Buffon (1707-1788) that New and Old World distributions of mammals were different (O'Malley, 2008). The geographic distribution of organisms was at the core of Darwin and Wallace's evolutionary programme based on the mechanism of natural selection.

Microbial geographical distribution has relied exclusively on ecological explanations under the fundamental assumption that *everything is everywhere: but the environment selects* (Baas Becking, 1934). This tenet, originally promulgated by the Dutch microbiologist Martinus Wilhelm Beijerinck (1851-1931) and articulated by Lourens G. M. Baas Becking (1895-1963) [2], does not mean that there are no biogeographical patterns, but due to microbial high dis-

[2] See Quispel (1998) for an overview of Baas Becking's life and work, and De Wit & Bouvier (2006) for a discussion about his original tenet.

persability and large population size, geography is usually thought to be irrelevant. The reasoning behind Beijerinck's statement was derived from Candolle (1778-1841), who pointed to exceptions (i.e. cosmopolitanism due to unrestricted dispersal) to Buffon's law of nature, and explained in *On the Origin of Species*:

> the seeds and eggs of many low forms being very minute and better fitted for distant transportation, probably accounts for a law which has long been observed, and which has lately been admirably discussed by Alph. de Candolle [3] in regard to plants, namely, that the lower any group of organisms is, the more widely it is apt to range. (Darwin, 1859, p. 406)

Beijerinck believed that microbiology could contribute to a universal theory of life (O'Malley, 2008). His convictions were transmitted to a broader audience by the emerging Delft School, and his successor Albert J. Kluyver (1888-1956) placed the basic principle at the core of general microbiology in his Leeuwenhoek Lecture to the Royal Society in London:

> In principle, owing to a continuous influx of airborne germs of all types, everything will be everywhere, but the ecological factors-the environment in the second part of Baas Becking's phrase-will decide which germs will proliferate, which germs will maintain themselves at a low numerical level, which germs will die off. Experience teaches us that many microbial species manage to survive at numerous spots on earth, these species may be called ubiquitous in a more strict sense of the word. On the other hand other species will only be found at places where very special conditions prevail. (Kluyver, 1953, p. 153)

The topic of microbial geographic distribution has been customarily simplified under Baas Becking's unchallenged tenet, but microorganisms dwell in diverse and complex communities whose intrinsic nature has been subject of intense debate. The nature of ecological communities has to be traced back to Stephen A. Forbes (1844-1930). He was impressed by the connectedness of plants and animals living together in a lake, and stressed *the impossibility of studying completely any form out of relation to the other forms; the necessity for taking a comprehensive survey of the whole as a condition to a satisfactory understanding of any part* (Forbes, 1887). Sergei N. Winogradsky (1856-1953), considered the first microbial ecologist[4], anticipated the ecosystem concept, and exemplified several principles of community structure in his studies of soil biology (Ackert, 2007). In general ecology, two opposed views structured the

[3] Darwin incorrectly cited the source as Alphonse de Candolle (the younger) instead of the elder, Augustin de Candolle (O'Malley, 2008).

[4] See Dworkin (2012) for a biographical sketch of Sergei Winogradsky.

debate during the 20th century: on the one hand a community as a group of interdependent and inextricably linked organisms (Clements, 1916), and on the other hand as random assemblages of individually distributed species (Gleason, 1926) supported by the distribution of populations over ecological gradients within regions (Whittaker, 1967). The gleasonian view has gained most support and recent definitions consider the community as an epiphenomenon with relatively little explanatory power (Ricklefs, 2008), or as a convenient assemblage at some arbitrarily study site for some arbitrary period of time that a scientist selects for study (Brown, 1995). Although macrobial ecologists have neglected any notion of communities with properties analogous to individual organisms, or superorganism, the perspective in microbial ecology may not be akin on behalf of the lack of clear boundaries (i.e. genetic exchange make boundaries more permeable; O'Malley & Dupré, 2007), and due to the existence of closed systems such as microbial mats that operate as individual entities with tightly coupled oxidation-reduction gradients and restricted vertical flows (Margalef, 1997; Guerrero et al., 2002).

One way to confront complexity is to adopt a holistic point of view (Margalef, 1963, 1968; Odum, 1969; Solé & Bascompte, 2006), and to increase the scale of description (Maurer, 1999). As local unpredictability is generally the most predictable feature of complex ecological systems, when the scale of description increases, variability declines and predictability augments (Levin, 1992). The discipline of macroecology (Brown, 1995) seeks to broaden the scope of ecology at much larger spatial and temporal scales by means of a comparative statistical methodology instead of an experimental manipulative one. Thus, it attains greater potential of generalization and synthesis but with a less detailed delineation of the phenomenon under study. As a consequence, *macroecology is as much about the deductive effort of evaluating hypotheses as the inductive effort of searching for patterns* (Brown, 1995, p. 232).

Microbial communities are the ideal system to study macroecology. The total number of prokaryotic cells is in the order of 10^{30}, eight orders of magnitude greater than the number of stars in the observable universe (Whitman et al., 1998), and the number of species may be in the order of 10^{6-9} (Curtis et al., 2002). Although biological systems are characteristically variable (i.e. no two biological systems are identical), patterns and regularities become evident at larger scales (Maurer, 1999). This statement may seem restricted to biology but *the laws of physics and chemistry are statistical throughout* [5] (Schrödinger,

[5] A basic assumption of quantum mechanics is that matter is composed of a large number of small particles with inherent indeterminate properties. Hence, regular laws of classical newtonian physics are only statistical approximations. The reason for regularity is that there are so many particles involved that uncertainty becomes insignificant. The uncertainty is on the order of $n^{-0.5}$, where n is the number of particles involved. For a litre of gas molecules, the error would be a

1944, p. 4). However, microbial communities are not completely sampled in practice. As an example, a clone library of size 100 from a community with actual size of 10^8 is only recovering one clone for every million individuals. Despite unavoidable incomplete sampling, sequencing effort does not appear to affect the ranking of richness (i.e. α-diversity; Shaw et al., 2008), or the patterns of community similarity (i.e. β-diversity; Kuczynski et al., 2010). Thus, the decreased cost of sequencing should be focused to increase the number of samples rather than a deeper coverage of samples.

2.2 Ecological patterns and processes, and the problem of scale

As enunciated by the ecologist Robert H. MacArthur [6](1930-1972), detecting patterns is the fundamental step in every scientific inquiry:

> The concept of pattern or regularity is central to science. Pattern implies sort of repetition, and in nature it is usually an imperfect repetition. The existence of the repetition means some prediction is possible -having witnessed an event once, we can partially predict its future course when it repeats itself. The imperfection of the repetition gives us the means of making comparisons. (MacArthur, 1972, p. 77)

Ecologists typically distinguish three elementary perspectives on the dominant factors that influence the patterns of community composition, diversity and assembly [7]. First, the classical deterministic niche-based following local environmental filters and the principle of competitive exclusion. Early naturalists, including Darwin, recognized that phenotypic attributes of species could influence their interactions with other species and with the environment in predictable ways. The idea that similar phenotypes should share habitat affinities was fundamental for the development of niche theory (Grinnell, 1924; Elton, 1946; Hutchinson, 1959, and reviewed in Chase & Leibold 2003). In the 1920s and 1930s, theoretical developments tried to integrate ecological and evolutionary thinking by the adoption of the principle of competitive exclusion [8] (Hutchinson, 1959, 1961; Hardin, 1960; MacArthur & Levins, 1967)

negligible 10^{-17}% (cited in Maurer 1999, p. 23, and original reference in Schrödinger 1944, p. 17).

[6]See Fretwell (1975) for a summary of MacArthur's influence on ecology.

[7]Alternatively, a recent synthesizing effort that mimics the framework of population genetics summarizes the processes influencing community ecology patterns to four classes: selection, drift, speciation, and dispersal (Vellend, 2010).

[8]Microbial experimental systems played a central role in the development of the principle of competitive exclusion (Jessup et al., 2004). G. F. Gause (1910-1986) coupled his laboratory experiments containing bacteria, yeast and protists with the mathematical models of Lotka and Volterra to ask *why has one species been victorious over another in the great battle of life* (Gause, 1934).

that led to the proposal of community assembly rules (Diamond, 1975). This basic idea around interactions limiting coexistence became one of the central paradigms of community ecology and reinforced the convenient assumption that evolutionary processes were not relevant at the time scales of ecological observed patterns (Cavender-Bares et al., 2009). Nevertheless, empirical studies and theoretical models indicated the presence of evolved trade-offs that prevent all species from occurring in all environments, thus permitting coexistence (Tilman, 1982; Chesson & Warner, 1981; Chesson, 2000).

Determinism (in the sense that if A causes B, then A must be always followed by B) does not entail complete predictability. Deterministic chaotic dynamics sensitive to initial conditions exist in ecological systems composed of complex networks of many species governed by multiple factors (May, 1974; Solé & Bascompte, 2006). Contrariwise, in a stochastic or random process there is some indeterminacy in its future evolution even though the initial conditions are fixed. Probability or lack of determinism does not impose absence of causation (i.e. A probabilistically causes B if A's occurrence increases the probability of B). Traditionally, stochasticity has been interpreted as imperfect knowledge of a deterministic system but as we move towards the present, the inherently indeterministic nature of causal systems is acknowledged [9]. Without entering into the philosophical debate between determinism and indeterminism (quantum mechanics have extended it to the scientific arena), ecological (and biological) systems are perceived with a certain degree of uncertainty (Mayr, 1961). Hence, the second perspective to the deterministic niche view postulates that community assembly is largely a neutral process [10] based on stochastic processes and dispersal. This position was explicitly rejected by Darwin due to his entire confidence on the determistic laws of nature:

> When we look at the plants and bushes clothing an entangled bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this! (Darwin, 1859, p. 74)

A century later, the idea of stochasticity played however a central role in the theory of island biogeography (MacArthur & Wilson, 1967), and gained new prominence with the unified neutral theory of biodiversity (Hubbell,

[9] Universe's deterministic nature is usually referred as the Laplace's demon. Pierre-Simon Laplace (1749-1827) stated that from the laws of classical mechanics and the knowledge of the precise location and momentum of every atom in the universe, the past and the future of the entire universe would be revealed.

[10] The idea of neutrality in population genetics was first introduced by Kimura (1968). The neutral theory of molecular evolution enunciates that the vast majority of evolutionary changes at the molecular level are induced by random drift of selectively neutral mutants. Both evolutionary and ecological embracement of stochasticity reflect the same movement that occurred in physics from classical Newtonian to relativistic Einstenean mechanics (Simberloff, 1980; Ulanowicz, 1999).

2001). This perspective considers communities as open, non-equilibrial assemblages of ecologically equivalent species, whose abundance are governed by random speciation and extinction, dispersal and ecological drift (Alonso et al., 2006). Curiously, most of the general community patterns (see Table 2.1 for a short list) have been proven to arise from mere neutral models (Bell, 2001). Although ecologists have tried to discern the relative influence of deterministic (niche) and stochastic (neutral) processes (for instance, by variance-partitioning; Legendre & Legendre, 1998); it may be difficult to detect a truly definable deterministic signal when there are lots of stochastic noise [11].

Finally, the third perspective emphasizes the role of historical factors (notably, speciation and former dispersal at the regional scale) over local processes (Ricklefs, 1987). Although this perspective has recently gained new interest, it can be traced back to Charles Lyell (1797-1875) who was aware that environmental determinism was insufficient to explain biogeographical regions or endemisms (Bueno-Hernández & Llorente-Bousquets, 2006) [12]. However, MacArthur (1965) argued that large-scale history and geography could be ignored in the study of ecological communities because local processes come into equilibrium so rapidly that processes on larger scales are inconsequential (see Ishida 2007 for a discussion about MacArthur's ahistorical approach to ecology). Although the increase of information may be an unavoidable manifestation of historical development (Margalef, 1997), historical contingencies [13] should be always considered when dealing with biological systems. Historical contingency is expected to be more relevant in ecosystems with large regional species pools and low colonization rates (Chase, 2003). Phylogenetic approaches (see section 2.4) may increase the ability to study historical events and thus, the ultimate causes of community organization

[11] Twenty-five centuries ago, Democritus already expressed this fundamental duality with his celebrated maxim *Everything existing in the universe is the fruit of chance and necessity* which gave name to Jacques Monod's book *Chance and Necessity* (Monod, 1971).

[12] Lyell's *Principles of Geology* strongly influenced Darwin's views on the geographical distribution of biological forms:

> We are thus brought to the question which has been largely discussed by naturalists, namely, whether species have been created at one or more points of the earth's surface. Undoubtedly there are very many cases of extreme difficulty, in understanding how the same species could possibly have migrated from some one point to the several distant and isolated points, where now found. Nevertheless the simplicity of the view that each species was first produced within a single region captivates the mind. He who rejects it, rejects the *vera causa* of ordinary generation with subsequent migration, and calls in the agency of a miracle. (Darwin, 1859, p. 352)

[13] For an excellent and popularizing account on the role of historical contingency in macroevolution see Stephen Jay Gould's book *Wonderful Life* (Gould, 1989).

and development because they can inquiry from proximate ecological circumstances to more evolutionary and/or biogeographical explanations (Losos, 1996).

Correspondingly, the main task of community ecologists is to disentangle the relative influence of niche-related, neutral and historical processes (see Figure 2.1) because ecological patterns arise through a intricate statistical process of causality. As mentioned before (see section 2.1), the classical view in microbiology stated that stochastic or historical events are not relevant. The recently developed theoretical framework of metacommunity ecology, that is both ecological and biogeographic, tries to accommodate most of these processes by explicitly including dispersal among communities (Leibold et al., 2004; Holyoak et al., 2005). Leibold et al. (2004) formulated four different perspectives: (i) species sorting, which assumes high dispersal and local determinism, (ii) mass effects occur when dispersal is so high that it results in source-sink dynamics, (iii) neutral, which assumes all species ecologically equivalent, and (iv) patch dynamics imply a trade-off between dispersal capacity and competitive ability. If microorganisms are highly dispersible, metacommunity structure is expected to be dominated by species sorting and mass effects rather than neutral or patch dynamics. Anyhow, local factors are in general more straightforward to measure, and historical events, such as past dispersal barriers or past environmental conditions, can only be presently observed as spatial effects (Lindström & Langenheder, 2012). Functional information (i.e. traits associated with the mechanisms by which organisms in a community interact with each other and their environment) is of capital importance for a deeper understanding of the interplay between patterns and processes. However gathering functional trait data for microbial communities is extremely difficult, as most microorganisms in nature defy cultivation (Whitman et al., 1998). As an alternative approach, metagenomic (community genomics) techniques allow significant proportions of environmental genomes to be sequenced (Handelsman, 2004).

As succinctly explored in this section, ecologists advocate for a plethora of processes underlying the observed community patterns (Figure 2.1). A supplementary layer of entanglement comes from the realization that some mechanisms operate at different levels, or a process is the origin of patterns at different levels (that is, a hierarchical structure [14]). Patterns and processes flourish one upon the other, and it becomes arduous to project the influence

[14]Stephen Jay Gould in his monumental *The Structure of Evolutionary Theory* advocated for a hierarchical interpretation of evolution based on six levels: gene, cell, organism, deme, species and clade (Gould, 2002). Darwin already considered evolution as a two-level structure: natural selection acting on individuals (i.e. microevolution), and differential proliferation of clades (i.e. macroevolution; for a discussion see Chapter 9 in Maurer, 1999)

of a process at a particular scale over remote patterns and processes. To understand any given level in a biological hierarchy, it is required to examine the properties of the lower level as well as the constraints (e.g. historical coercions) imposed by the higher level (Salthe, 1985). No single mechanism explains trends on all scales, and the scale of observation certainly influences the description of pattern (Levin, 1992). As a general rule, when increasing the scale, the structure and behaviour of complex systems become predictable (i.e. emergent patterns).

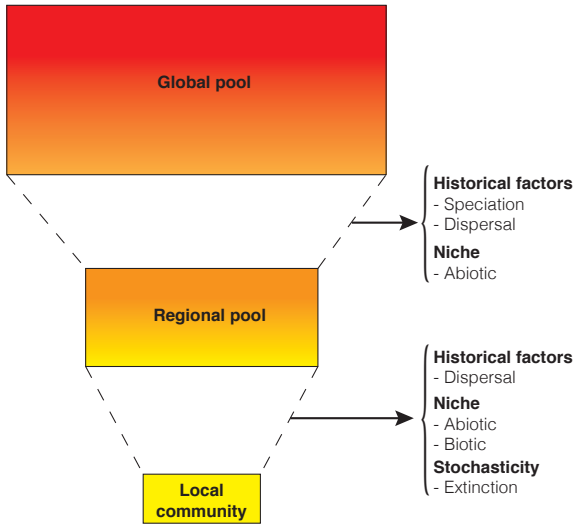


Figure 2.1: Conceptual summary of the main processes influencing community composition, structure and diversity at different spatial scales. All ecological and evolutionary processes considered have been encapsulated in three perspectives: deterministic (i.e. the biotic and abiotic niche), stochastic, and historic. The demarcation of discrete spatial scales is arbitrary. Figure modified from Soininen (2012).

Community assembly operates on both ecological and evolutionary time scales, resulting in both recent and historical elements, and spatial patterns arise by means of disparate regional and local processes (Hillebrand & Blenckner, 2002). Accordingly, it seems very difficult to link short-term local processes to global processes that occur over evolutionary time scales. At local geographic scales with no dispersal limitation, environmental heterogeneity is expected to be the major driver, while across larger scales the effects of dispersal limitation become more relevant and endemic variants may arise (see a recent review focused on microorganisms in Whitaker, 2006). The problematic issue with scale has resulted in the traditional separation between the fields of

ecology and biogeography (Wiens & Donoghue, 2004).

Besides the temporal and spatial scale, there also exists the source of hierarchy imposed by evolution on life forms (i.e. the taxonomic scale [15]). Studies of large clades are less likely to be well-resolved and will not yield insight into short-term patterns, but instead provide information about regional speciation, extinction, and biogegraphical patterns from older time periods. We would expect that the influence of historical contingency or dispersal limitation may be more apparent both at finer levels of taxonomic resolution and at the global scale (Fierer, 2008). Moreover, there exists hierarchical variability regarding guild or trophic levels. For instance, nearly all plants share the same resources while the metabolic diversity of microorganisms is astonishing (Kluyver & Van Niel, 1956).

2.3 Idiosyncrasy of microbial communities

Microorganisms are recognized as key players in Earth's ecosystems (Falkowski et al., 2008), but the diversity and ecology of most microbial assemblages remains poorly understood (Curtis et al., 2006). A more complete understanding of microbial patterns and processes is essential in order to interpret ecosystem functions and to predict the Earth's response to global change (Fuhrman, 2009). However, only very recently microbial ecologists have begun to inquiry about the patterns long known in the ecology of plants and animals (Table 2.1).

The small size of microorganisms is of vital importance to comprehend ecological scaling in structured assemblages. As stated by O'Malley & Dupré (2007), an excesive focus on macroorganisms may have distorted several basic aspects of our view of life. D'Arcy Thompson in his classical book *On Growth and Form* early recognized the size scale dependence in the interpretation of the world and how the microbial world scapes our regular human intuition:

> In the end we begin to see that there are discontinuities in the scale, defining phases in which different forces predominate and different conditions prevail. Life has a range of magnitude narrow indeed compared to that with which physical science deals; but it is wide enough to include three such discrepant conditions as those in which a man, an insect and a bacillus have their being and play their several roles. Man is ruled by gravitation, and rests on mother earth. A water-beetle finds the surface of a pool a matter of life and death, a perilous entanglement or an indispensable support. In a third world, where the bacillus lives, gravitation is forgotten,

[15] For instance, the domain *Bacteria* is estimated to be approximately 3.5 billion years old (Schopf & Packer, 1987), more than thirty times older than the ancestor of all birds (Padian & Chiappe, 1998).

Table 2.1: Some classical ecological patterns investigated using microorganisms. Note the time lag between the original references and the microbial examples. See recent reviews in Martiny et al. (2006); Dolan (2006); Ramette & Tiedje (2007); Fierer (2008); Soininen (2012).

Pattern	Explanation	Original references	Microbial examples
Species abundance distribution	Although microorganisms comprise many rare species, the pattern does not deviate from larger organisms (i.e. log-series and log-normal distributions).	Fisher et al. (1943); Preston (1948)	Curtis et al. (2002)
Species-area relationship	Weaker but significant power-law curve for microorganisms.	Arrhenius (1921); Gleason (1922)	Horner-Devine et al. (2004); Reche et al. (2005)
Distribution-abundance relationship	Microorganisms show similar strong relationships between regional distribution and local abundance.	Brown (1984)	Soininen et al. (2011)
Distance-decay relationship	Microorganisms tend to show moderate significant slopes with low initial similarity.	Willis (1922); Cain (1944)	Green et al. (2004)
Latitudinal gradient	Observed in marine microbial communities, but not in soils.	Darlington Jr (1959); Fischer (1960) [a]	Fierer & Jackson (2006); Pommier et al. (2007); Fuhrman et al. (2008)
Elevational gradient	Unicellular taxa radically differ from larger plants and animals.	Orians (1969); MacArthur (1972)	Bryant et al. (2008); Fierer et al. (2011)
Diversity-productivity relationship	*Bacteria* resembles larger organisms, but different groups show different responses.	Connell & Orias (1964); Leigh Jr (1965)	Horner-Devine et al. (2003)

[a] Latitudinal gradients were recognized by early naturalists more than a century ago.

> and the viscosity of the liquid, the resistance defined by Stokes's law, the molecular shocks of the Brownian movement, doubtless also the electric charges of the ionised medium, make up the physical environment and have their potent and immediate influence on the organism. The predominant factors are no longer those of our scale; we have come to the edge of a world of which we have no experience, and where all our preconceptions must be recast. (Thompson, 1942, p. 771)

A key yet unsolved question is whether microbial community patterns differ from those of macroscopic organisms just because of size, or they are in fact radically different from the rules affecting plants and animals and thus, they cannot be extrapolated to microorganisms. Regarding this, Brown (1995) postulated that:

> The physiological and ecological constraints of body size and the fundamental features of population and species dynamics are common to all organisms. Consequently, I suspect that many of the macroecological patterns and processes, with only minor or quantitative modifications, will be equally general. (Brown, 1995, p. 174)

Anyhow, the biology of microorganisms entails some particularities that deserve further attention for applying a community ecology approach. First of all is the species concept. The genetically isolated lineage, often conceived as the fundamental unit of evolution, may have no real analogue in the microbial asexual world, and hence most of life and its history cannot be simply conceived as an intelligible tree-like pattern (Maynard Smith et al., 1993; Doolittle & Zhaxybayeva, 2009). Microorganisms refuse to show definable and consistent taxonomic classification schemes, and it is unclear if they should be treated as continua or as discrete clusters (Rosselló-Mora & Amann, 2001; Konstantinidis & Tiedje, 2007). If horizontal gene transfer is far less common than mutation, the situation is essentially clonal with a high degree of clustering (Fraser et al., 2007). This seems to be the actual case with an estimated gene uptake rate by lateral transfer on the order of 10^{-4} to 10^{-6} (Fraser et al., 2007) and a mutation rate of 0.0033 per genome per DNA replication (Drake, 1991). Empirically, a definition of bacterial species has been based on DNA-DNA cross-hybridization or sequence similarity in the 16S rRNA gene. Less than 70% hybridization or less than 97% similarity have been proposed to belong to different species [16] (Stackebrandt & Goebel, 1994). As an alternative for asexual microorganisms, the ecological species concept defines a species as a set of individuals showing genetic cohesion and identical in their ecological properties (ecotypes), and predicts that genetic diversity directly relates to

[16] Application of the same definition to apes would classify humans, orangutans and gibbons as the same species (Sibley & Ahlquist, 1987).

ecological diversity (Cohan & Koeppel, 2008). Microbial evolution can occur very rapidly (compared to macroorganisms), potentially leading to convergence of ecological and evolutionary time scales (Sniegowski et al., 1997). It has been proposed that the large number of microbial species is due to low extinction and high speciation rates (Dykhuizen, 1998).

Dispersal is a key mechanism for understanding community assembly and structure because it can operate in short-term ecological scale and as historical contingency. The potential extent of microbial dispersal is expected to be huge (Bovallius et al., 1978), although successful colonization actually involves arrival and establishment. However, asexual microorganisms tend to be excellent colonizers because of their dispersal capabilities and because a single individual can form a new local population (Brown, 1995). The proportion of species that are ubiquitous [17] purportedly decreases abruptly at 1 mm, and cosmopolitanism of *Bacteria*, *Archaea* and protists genera is higher than that observed for macroorganisms (Finlay, 2002). Highly dispersive species are less likely to show a signature of their biogeographic history and are expected to show more effects of local environment and less effects of proximity in their β-diversity patterns (e.g. Beisner et al., 2006). However, specialized microorganisms to rare or extreme habitats may experience dispersal limitation and geographic barriers (e.g. Whitaker et al., 2003), though measures of actual microbial dispersal rates are limited.

A key question about the influence of dispersal is when (or at which scales) does colonization or *in situ* evolution predominate in the assembly process. Communities assembled through *in situ* evolution vs. those assembled through dispersal may exist as a dynamic equilibrium between the two extremes of a continuum (Curtis et al., 2006; Cavender-Bares et al., 2009) because available ecological space is filled either by adaptation of early occupants or by foreign colonization, depending on which occurs first. The observation that ecologically relevant traits are phylogenetically conserved has lent support to the hypothesis that it is more feasible to move than to evolve (Cavender-Bares et al., 2009). Certainly, potential links between community ecology and macroevolution exist, though the biggest challenge is to reconcile the mismatches of evolutionary and ecological patterns that emerge at different spatial and temporal scales. Regretfully, the fossil record which is the richest source of information on the historical events behind extant communities (Jablonski & Sepkoski Jr, 1996) is mostly absent for *Bacteria* and *Archaea* (see some exceptions in Schopf & Packer, 1987; Coolen & Overmann, 2007).

[17] The most ubiquitous organism found to date is the alphaproteobacterial SAR11 clade, predicted to comprise more than 20% of the total cells in the marine environment (Morris et al., 2002).

Besides dispersability, high speciation or low extinction, it could be considered dormancy as the crucial microbial distinctiveness for the comprehension of the observed community patterns. Dormancy (i.e. a reversible state of low metabolic activity) is a common strategy that allows microbial species to contend to temporal variability of environmental conditions. The proportion of dormant cells may represent up to 50% of total microbial cell counts in natural communities (Lennon & Jones, 2011). Dispersal and dormancy are alternative ways to reduce the experienced environmental variability (Levin, 1992). High dispersal and dormancy are expected to be fundamental for the concept of the rare biosphere illustrated by microbial rank abundance curves with extremely long tails (i.e. dominated by low-abundance taxa). This concept is analogous to the seed bank that represents a reservoir of genetic diversity that is capable of responding to environmental change, contributing to the diversity and dynamics of future generations (Pedrós-Alió, 2006; Jones & Lennon, 2010). It also allows competing species to coexist via the storage effect (Chesson & Warner, 1981; Chesson, 2000). Overall, dormancy should reduce the risk of local extinction through the recruitment from the seed bank, and should increase the probability of successful colonization by avoiding mortality (Figure 2.2). In the absence of dormancy, persistence is more dependent on immigration. Additionally, the effects of dormancy should increase the expected richness in a given habitat (α-diversity) but decrease species turnover (β-diversity; Locey, 2010; Lennon & Jones, 2011, and see Figure 2.2), and should enhance species sorting (see previous section). Short generation times and dormant seed banks may result in strong numerical advantage of first colonizers (i.e. priority effects and monopolization) in microbial habitats. Serial colonization may yield a pattern of isolation by distance not driven by space itself but because of historical colonization events, and thus, mosaic and enclave distributions may be much more common in microorganisms than in macroorganisms (De Meester, 2011).

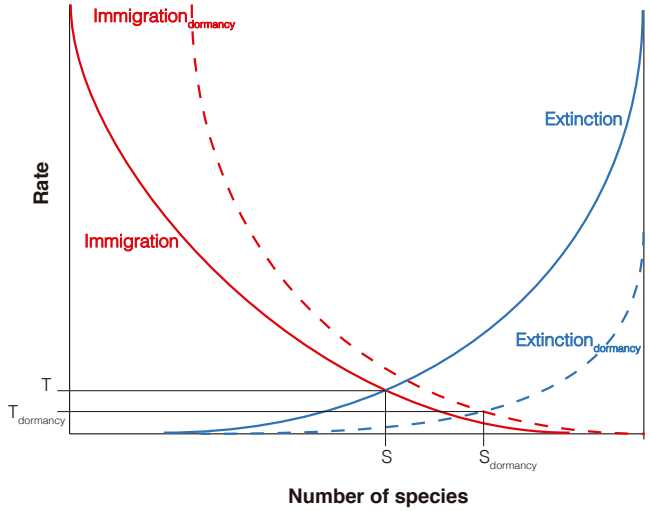


Figure 2.2: Graphical representation of the predictions of the theory of island biogeography. According to MacArthur & Wilson (1967), the number of species (richness) on an island depends primarily on the relative rates of immigration from the mainland and of extinction on the island. If so, richness will eventually reach an equilibrium when the rate of immigration is balanced by the rate of extinction. Thus, $S = \frac{IP}{E+I}$, where *S* is the equilibrium richness, *E* and *I* the rates of extinction and immigration respectively, and *P* the size of the source pool from the mainland. Dormancy increases immigration rates and decreases extinction rates. As a consequence, richness is higher ($S_{dormancy}$ > S), but turnover is lower ($T_{dormancy}$ < T). Figure modified from Locey (2010).

2.4 Phylogenetic ecology as an integrative framework

As summarized in previous sections, biological (and ecological) systems are the result of a number of deterministic and stochastic processes as well as historical constraints. The most relevant historical coercion is that species are not independent entities, but their functional and ecological similarities are rather shaped by patterns of common ancestry:

> Species are part of a hierarchically structured phylogeny, and thus cannot be regarded for statistical purposes as if drawn from the same distribution. (Felsenstein, 1985, p. 1)

Or in Darwin's words:

> In considering the Origin of Species, it is quite conceivable that a naturalist, reflecting on the mutual affinities of organic beings, on their embryological relations, their geographical distribution, geological succession,

> and other such facts, might come to the conclusion that each species had not been independently created, but had descended, like varieties, from other species. (Darwin, 1859, p. 3)

Phylogenetic comparative methods pioneered by the method of phylogenetic independent contrast (PIC) meant the first substantial effort to address the statistical non-independence among species due to common ancestry (Felsenstein, 1985; Ackerly, 2009). In the field of community ecology, only recently researchers have incorporated historical constraints represented as phylogenies in their analyses motivated by the fact that species interact within the community based on their traits, and traits have an evolutionary history (Webb et al., 2002). In a hypothetical world in which evolution was rapid and unconstrained and in which any lineage had no dispersal limitation, communities in similar environments would be also similar. However, in the real world, evolution is constrained and lineages tend to be restricted in geographic distribution (Losos, 1996). In order to account for this evidence, a set of tools that aim to bridge the gap between evolutionary and ecological analyses have been recently developed. Nowadays, ecologists can assess: (i) where most of the biological diversity accumulates (Faith, 1992) and how it is intrinsically structured (Webb, 2000; Helmus et al., 2007), and (ii) how phylogenetic β-diversity (i.e. similarity among communities based on evolutionary history) is distributed along environmental gradients (Lozupone & Knight, 2005; Bryant et al., 2008; Ives & Helmus, 2010).

Species richness alone does not provide a measure of the diversity of lineages that live in a region. Phylogenetic diversity (PD), the sum of the branch lengths from the members within a community, is more inclusive that a simple count of species richness because it quantifies the evolutionary history, and is also believed to correspond to the number of evolutionary derived traits within a biological community (Faith, 1992). Hence, PD will be higher when there are more distantly related species in an assemblage.

Community phylogenetic structure [18] (i.e. the degree of phylogenetic relatedness within a community) provides insight into the ecological and evolutionary drivers of community assembly. For instance, if closely related species are similar in resource use, and competitive exclusion prevents co-occurrence,

[18] There exist two types of metrics of community phylogenetic structure. On the one hand, the mean pairwise distance (MPD) measures the mean phylogenetic distance between all members from each community, and the mean nearest taxon distance (MNTD) calculates the mean distance separating each member from its closest relative (Webb, 2000; Webb et al., 2002). On the other hand, the phylogenetic species variability index (PSV) quantifies how phylogenetic relatedness decreases the variance of a hypothetical neutral trait shared by all members of a community. The value is 1 when all species are unrelated (i.e. a star phylogeny) and approaches 0 as species become more related (Helmus et al., 2007).

then one expects that close relatives will not tend to coexist within local communities, or quoting Darwin:

> As species of the same genus have usually, though by no means invariably, some similarity in habits and constitution, and always in structure, the struggle will generally be more severe between species of the same genus, when they come into competition with each other, than between species of distinct genera. (Darwin, 1859, p. 76)

However, if environmental filtering is prevalent, close relatives that share adaptations to particular habitats will co-occur more than expected. Because many biological traits are generally conserved during the evolution of a lineage, one would expect a positive relationship between phylogenetic relatedness and ecological similarity; that is, on average species resemble their close relatives (Losos, 2008). Phylogenetic clustering is consistent with environmental filtering or determinism, while phylogenetic overdispersion or evenness can be explained by multiple factors such as biotic interactions (i.e. competition or facilitation), small-scale habitat heterogeneity, and even a filtering process when the traits that promote success have evolved independently (Webb et al., 2002; Cavender-Bares et al., 2009). Measuring traits is essential to differentiate between processes because phylogenetic community analysis in the absence of trait data has limited insights (Vamosi et al., 2009). Moreover, the challenge is to explicitly define the ecological, spatial, and taxonomic scales (see section 2.2) because the processes that structure the assembly of regions, communities, and habitats differ, and can confound inferences about ecological and evolutionary processes (Webb et al., 2002). Results available so far suggest a stronger role for resource partitioning among closer relatives and at smaller spatial scales, whereas habitat filtering dominates at larger spatial and phylogenetic scales (see recent reviews in Vamosi et al., 2009; Cavender-Bares et al., 2009).

Phylogenetic turnover or β-diversity or community similarity can be defined as the fraction of branch length shared between two communities [19] (Lozupone & Knight, 2005; Bryant et al., 2008). Taxonomic or compositional β-diversity and phylogenetic β-diversity would be exactly the same

[19] Specifically, let A denote the sum of branch lengths that are shared between members of communities x and y; let B denote the sum of branch lengths (that is, PD) leading only to members of community x; and let C be defined correspondingly but for community y. Then,

$$UniFrac_{x,y} = \frac{B + C}{A + B + C} \qquad \text{(Lozupone \& Knight 2005)}$$

$$PhyloSor_{x,y} = \frac{B + C}{2A + B + C} \qquad \text{(Bryant et al. 2008)}$$

between two communities if every species were equally related to every other one, that is, a star phylogeny.

As constantly stated by Ramon Margalef [20], ecology (i.e., the interplay between biodiversity and the environment) is the scenario where evolution (i.e., the historical component of biological change) plays (Margalef, 1963). Phylogenies derived from molecular data provide an indirect record of the speciation events that have led to extant species, reflecting the tempo and mode of macroevolutionary processes related to diversification (Mooers & Heard, 1997). However, the study of macroevolution in microorganisms suffers from two obvious flaws: the lack of fossil records and the unknown scope of the true diversity (Curtis et al., 2002). Different interpretations have arisen when explaining the observed patterns of microbial diversification along time. In a recent meta-analysis using phylogenetic trees inferred for microorganisms, the general pattern conformed to expectations assuming a constant speciation rate (Martin et al., 2004). Contrastingly, a meta-analysis using molecular phylogenies of macroorganisms indicated that although clades showed great heterogeneity, most of them showed rapid lineage accumulation early with a slowing more recently. The missing component in these macroevolutionary interpretations is the ecological context of speciation and extinction. Clades diversify in an ecological context, but most models do not directly encapsulate the ecological mechanisms that influence speciation. A recently published metacommunity model was able to generate the full range of patterns by simply manipulating the degree of ecological differentiation of new species at the time of speciation (McPeek, 2008). The model implies that for a complete understanding of the true macroevolutionary dynamics of ecosystems, the ecological interactions that shaped the history of the clades have to be accounted for.

By providing a temporal dimension to community ecology, phylogenetic information allows ecologists to assess when and where traits originated, and consequently, whether communities are primarily assembled through *in situ* evolution or through dispersal (Cavender-Bares et al., 2009). Thus, incorporating phylogenetic information into community ecology is enlightening because

[20] Ramon Margalef (1919-2004) was the pioneer of ecology in Spain (for a short introduction to his life and work see Herrera, 2005). His intuitive and integrative approach to nature was summarized by himself:

> For most topics that concern ecology, I like poets more than lawyers, and feel more inclined to phantasy, feeling and inspiration than to rigor, consistency and even responsibility. In my views on environmental problems, I feel more attracted by the origin of the troubles and what they tell us about the workings of the biosphere than by their solutions, at least in the way the problem is usually being faced at present. (Margalef, 1997, p. 2)

it allows ecological questions to be addressed in an evolutionary context:

> Phylogenetic approaches thus should be seen as an integral component of studies of the causal basis of community structure [...] if one is interested in comparing the structure of communities or investigating how a community came to its present state, then phylogenetic information is indispensable for both generating and testing hypotheses. (Losos, 1996, p. 1352)

For uncultured microorganisms, how they differ in the genetic or physiological traits used to exploit the environmental resources is largely unknown, but we can start exploring the ecological processes underlying community assembly and diversification patterns through a phylogenetic ecology framework that takes into account the non-independence of living organisms because *the living system performs the role of a witness of the historical progress* (Margalef, 1996).

3
Objectives

There are no better words to commence this section than those of the polymath scholar Sir D'Arcy Wentworth Thompson (1860-1948):

> Natural history deals with ephemeral and accidental, not eternal nor universal things; their causes and effects thrust themselves on our curiosity, and become the ultimate relations to which our contemplation extends. (Thompson, 1942, p. 3)

Coarsely mimicking the beautiful prose of D'Arcy Thompson, I can say that my contemplation during the PhD period has extended on ephemeral and accidental microbial communities, along with their effects (i.e. patterns) and causes (i.e. processes).

Precisely, the general $\tau\acute{\epsilon}\lambda o\varsigma$ of any PhD dissertation should be explicit in its title. Hence, in this thesis I argue that novel insights into microbial community patterns arise when phylogenetic relatedness are used in conjunction with multivariate statistical techniques in the context of broad scales of description. The multivariate statistical technique of ordination was originally developed in ecology to array species along synthetic statistical gradients that represented covarying associations without presupposing any particular process (Bray & Curtis, 1957; Whittaker, 1967). With the additional information of environmental variables (as a surrogate of abiotic determinism) and spatial distribution (as a surrogate of an admixture of historic and stochastic events), I have been able to explore some possible mechanisms underlying the structure and diversity of microbial communities (Legendre & Legendre, 1998). In order to address the problem of non-independence of species due to common ancestry (Felsenstein, 1985), I made extensive use of phylogenetic techniques

whenever it was appropriate (Webb et al., 2002). The detailed objectives following the structure of this dissertation are given below:

Part I: Phylogenetic meta-analysis [1] of microbial 16S rRNA sequences at a global scale

The composition and diversity of microbial communities can now be assessed with molecular approaches, even when most of the members have never been cultured so little is known about their functional or ecological attributes. All four chapters included under this section share a global scale meta-analytical motive because *general principles can only come from comparative analyses of many studies done using comparable techniques of data collection and analysis or by expanding the scale of data collection* (Maurer, 1999, p. 54).

The specific objective of the first two chapters was to contrast different water masses (i.e. the oceanic continuum *versus* lakes as islands embedded in a sea of land, and the epilimnia *versus* the isolated hypolimnia of stratified lakes). Aquatic environments are ideal ecosystems to test differences in bacterial community composition and structure caused by dispersal limitation and environmental filtering.

- **Chapter 4** compares the bacterioplankton composition, community phylogenetic structure and phylogenetic β-diversity patterns between lakes and oceans, and for different taxonomic groups (Barberán & Casamayor, 2010).
- **Chapter 5** compares planktonic freshwater bacterial diversity and phylogenetic β-diversity patterns between the well-oxygenated upper water mass (epilimnion) and the anoxic bottom water compartment (hypolimnion) from stratified lakes (Barberán & Casamayor, 2011).

The last two chapters of this section are devoted to the phylogenetic ecology of uncultured *Archaea*.

- **Chapter 6** uses a global approach similar to Lozupone & Knight (2007) in order to assess the environmental factors that shape archaeal diversity, phylogenetic community structure and taxa distribution across

[1] Rigorously, the statistical term *meta-analysis* is not correctly used. Instead I should have described my approach as *comparative analysis*. However, this term applied to the context of phylogenetic information appeals to a set of methods, inaugurated by the evolutionary biologist Joseph Felsenstein, with a totally different purpose (Felsenstein, 1985).

seven distinct natural habitats (freshwater plankton, freshwater sediment, soil, marine plankton, marine sediment, hypersaline plankton, and hydrothermal vents; Auguet et al., 2010).

- **Chapter 7** focuses on two diverse groups of freshwater *Euryarchaeota*, Lake Dagow Sediment (LDS) and Rice Cluster-V (RC-V). We explored the ecological differentiation along the evolutionary history of both groups, and their temporal diversification patterns (Barberán et al., 2011).

Part II: Bacterial community ecology in Pyrenean lakes and the influence of Saharan desert dust deposition

Once stated the viability of the phylogenetic approach used in **Part I** and the network analysis developed in **Chapter 10**, the focus moved from comparative analyses of microbial communities at the global scale to the analysis of original data at the regional scale. The ecosystem chosen was the high mountain lakes located in the Spanish Pyrenees mountain range within and around the AigüesTortes and Estany de St. Maurici National Park. The study of this system began in the early 80s with the initial impulse of Ramon Margalef (Catalan et al., 2006). The change of scales allowed the investigation of the impact of global processes (such as atmospheric dust deposition across continents) in microbial assemblages at the local and regional scales.

- **Chapter 8** describes the bacterial community composition of eighteen high-mountain lakes in the Pyrenees, compares their phylogenetic diversity with other aquatic environments, and identifies the main environmental and geographical factors driving β-diversity patterns. High mountain lakes conform an interesting biogeographical model of spatially close but environmentally heterogeneous ecosystem. In this study, we found a significant relationship between lake area and phylogenetic diversity (Manuscript in preparation).

- **Chapter 9** applies pyrosequencing techniques to characterize the microbial communities inhabiting three different habitats associated with airborne dust: the source of dust plumes (i.e. soil samples from the Saharan desert), the atmospheric carrier (i.e. dust deposition sinking on the Central Pyrenees area), and the collector of the deposited dust (i.e.

the neuston of high mountain lakes). Although the short reads generated by high-throughput sequencing techniques circumvent the inference of accurate phylogenies, in collaboration with colleagues from the University of Colorado, we tried to determine the potential for dust-associated microbial communities to successfully colonize distant environments. Hence, we explored how microbial diversity at the regional scale is linked by long-distance dispersal to the global scale (Manuscript in preparation).

Part III: Novel ecological approaches to pyrosequencing and metagenomics data

Recently, the study of microbial community ecology has advanced considerably due to methodological advances such as pyrosequencing (Sogin et al., 2006) and metagenomics (Venter et al., 2004) that generate copious molecular information. However this rate of information gathering is exceeding the ecological interpretation rate. In this section, I applied analytical tools from complex systems science (i.e. network analysis) and general ecology (i.e. functional traits analysis) in order to confer ecological meaning to large datasets of genetic information. Thus, the main objective of this section was to bring methods and concepts from other fields to microbial ecology.

- **Chapter 10** explores co-occurrence patterns of soil microorganisms by network analysis. Thus, we tried to analyze the biotic component by itself without explicitly incorporating the environmental or spatial components (although it is implicitly assumed in the correlation approach used). We used a large pyrosequencing dataset obtained by colleagues from the University of Colorado that consisted of more than 160,000 bacterial and archaeal 16S rRNA sequences from 151 soil samples distributed worldwide (Barberán et al., 2012a).

- **Chapter 11** explores a set of community traits in 53 metagenomic aquatic samples from the Global Ocean Sampling expedition (Rusch et al., 2007). In collaboration with colleagues from the University of Oregon, we propose that some measures summarize the properties of the community as a single unit and may help to interpret complex metagenomics data. Furthermore, we compared the taxonomic and functional contents of microbial communities with the objective of trying to solve whether local assemblages adapt more efficiently by the modification of the genomic repertoire (i.e. functional adaptation) or by the taxonomic re-

placement (i.e. dispersal and colonization) of their members (Barberán et al., 2012b).

Part I: Phylogenetic meta-analysis of microbial 16S rRNA sequences at a global scale

4

Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities

Resumen

Este trabajo intenta identificar patrones filogenéticos de comunidades de grupos de bacterias abundantes en el plancton (*Alpha-*, *Beta-*, y *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, y *Bacteroidetes*) en sistemas acuáticos superficiales distribuidos globalmente (34 localizaciones -tanto lagos como mares- que suman 4500 secuencias del gen 16S rRNA). En cada localización se estimó la riqueza de grupos bacterianos y sus relaciones genéticas mediante una aproximación de filogenia de comunidades con el fin de (i) explorar aquellos procesos ecológicos compatibles con los patrones filogenéticos observados, y (ii) desentrañar los efectos espaciales y del ambiente en los patrones de β-diversidad para los diferentes grupos de bacterias.

Las masas acuáticas terrestres presentaron significativamente una mayor riqueza y diversidad de grupos bacterianos que las localizaciones marinas. Éstas, por su parte, presentaron un porcentaje mayor de emplazamientos agre-

gados filogenéticamente. Los análisis de β-diversidad revelaron pautas contrastadas en función tanto de la composición de sales (aguas marinas/masas acuáticas terrestres) como de la concentración salina. Se observó que dichos patrones para el grupo *Bacteroidetes* estaban estructurados según la concentración salina, mientras que en *Alphaproteobacteria* y *Gammaproteobacteria* estaban controlados según la composición de sales. Los grupos de *Actinobacteria*, *Betaproteobacteria* y *Sphingobacteria* fueron raramente detectados en aguas marinas y aguas continentales salinas. En general y pese a la falta de una profusión de datos contextuales, la similaridad ambiental fue más determinante que la espacial para la distribución de la β-diversidad de bacterioplancton. No obstante, se detectó una señal geográfica para algunos grupos (i.e. *Actinobacteria*, *Beta-*, y *Gammaproteobacteria*) en aguas continentales. A grandes rasgos, los análisis indicaron diferencias entre grupos filogenéticos y reflejaron patrones interesantes para investigaciones subsiguientes en torno a la formación de comunidades microbianas.

Abstract [1]

We aimed to identify phylogenetic community patterns in abundant planktonic bacteria (*Alpha-*, *Beta-*, and *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, and *Bacteroidetes*) from a worldwide range of surface waters (lakes and seas -34 sites and ca. 4,500 16S rRNA gene sequences). At each site we assessed the number of observed bacterial groups and the genetic relatedness of the most abundant groups through a community phylogenetic meta-analysis approach in order to (1) explore which potential ecological processes were consistent with the observed phylogenetic patterns in community assembly and (2) disentangle the effects of space and environment in β-diversity patterns for the different bacterial groups. Inland waters had significantly more bacterial groups and were more diverse than marine waters. Marine habitats showed a higher percentage of clustered sites than lakes, and bacterial communities were more closely related than expected by chance. Phylogenetic β-diversity analyses revealed different patterns to both salt composition (marine vs. inland salt lakes) and salt concentration for the dominant bacteria. We observed that while β-diversity patterns for *Bacteroidetes* were mostly shaped by salinity concentration, patterns in *Alphaproteobacteria* and *Gammaproteobacteria* were controlled by salt composition. *Actinobacteria*, *Betaproteobacteria* and *Sphingobacteria* were largely absent from marine habitats and from saline continental sites. In general and despite the lack of contextual metadata, environmental similarity was more relevant than spatial distribution for bacterial β-diversity patterns. However, we detected a geographic signal for some inland waters' groups (i.e. *Actinobacteria*, *Beta-*, and *Gammaproteobacteria*). Overall, the analyses indicated differences among phylogenetic groups and reflected patterns upon which further exploration of community assembly theory could be based.

4.1 Introduction

Biogeography explores how and why biological diversity changes along geographical scales. In the case of microorganisms, this topic had until recently received very little attention; the traditional statement "everything is everywhere, but the environment selects" (Baas Becking, 1934) summarized what was expected for microorganisms, i.e. on the one hand, high dispersal rates due to large population sizes and short generation times (Fenchel, 2003) and, on the other hand, environmental determinism in agreement with classical niche-assembly theories. Nonetheless, endemism, dispersal limita-

[1] See original publication in Barberán & Casamayor (2010).

tion and stochasticity have also been recently shown in different microbial groups (e.g. Whitaker et al., 2003). In this context, the timely debate between neutral and niche views on species distribution offers many opportunities to test new hypotheses with microbial communities. The neutral theory of biodiversity (Hubbell, 2001) assumes that the abundance of species is driven by random dispersal and stochastic extinction, and species are assumed to be ecologically identical. Conversely, the niche-assembly view looks at environmental adaptations to explain the abundance and distribution of species (Hutchinson, 1961).

The emerging field of metacommunity ecology tries to combine the neutral and niche views, recognizing different dispersal and niche-based processes (see a recent review by Leibold et al., 2004). Metacommunity ecology focuses on sets of local communities linked by the potential dispersal of individuals. Two opposing forces potentially shape local community structure: different local environmental factors could lead to divergence of the communities, while high dispersal rates could homogenize the connected communities. Different types of metacommunities have been proposed depending on the relative importance of dispersal and niche-based processes (Cottenie, 2005), and each type conceptualizes different complex dynamics of potential importance for the metacommunity. The neutral model (NM) assumes that random processes drive species abundances. Conversely, the species-sorting (SS) view emphasizes environmental adaptations to explain the abundance and distribution of species. Under the patch-dynamics model (PD), community composition is defined by a tradeoff between dispersal and competitive ability, with no habitat quality differences among patches. Finally, the mass-effect model (ME) mainly relies on the effect of immigration on local dynamics. In the case of aquatic ecosystems, these concepts have been applied to zooplankton (e.g. Cottenie & De Meester, 2004) and recently to bacterioplankton community ecology (e.g. Lindström & Logue, 2008).

For bacteria, many worldwide comparable 16S rRNA gene surveys are accessible from public databases, facilitating the integration of phylogeny and community ecology. Initial attempts have focused on the global distribution of prokaryotic communities in oceans (Pommier et al., 2007; Fuhrman et al., 2008), soils (Fierer & Jackson, 2006), and a mixture of terrestrial and aquatic environments (Lozupone & Knight, 2007; Auguet et al., 2010), and general ecological patterns have emerged from the study of 16S rRNA gene datasets (e.g. taxa-area relationships; Horner-Devine et al., 2004; Reche et al., 2005). Statistical methods that compare phylogenetic tree topology of different communities and metrics that quantify the distribution of taxa in a single sample relative to a pool of taxa (Webb, 2000) can be applied to microbes (Horner-Devine & Bohannan, 2006; Newton et al., 2007).

Here we focused on whether or not phylogenetic community patterns exist for some of the most abundant planktonic bacterial groups (*Alpha-*, *Beta-* and *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria* and *Bacteroidetes*) in a worldwide range of lakes and seas. Even though the bacterial groups selected have members with different ecophysiology, consistent ecological trends have been observed at such broad clustering levels (e.g. Glöckner et al., 1999; Bouvier & del Giorgio, 2002; Demergasso et al., 2004). At each site, we assessed the community phylogenetic diversity with an index (Faith, 1992) that takes into account the number of sequences (i.e. species richness) and the phylogenetic relationship among them (branch length). The different tree topologies (shape) were examined using two indices, i.e. the net relatedness index (NRI) and the nearest taxa index (NTI; Webb, 2000) to estimate whether or not the bacterial assemblages exhibited significant phylogenetic structure. Although different scenarios could lead to the same patterns observed in the data, it has been proposed that those communities primarily structured by competitive exclusion would be less closely related than expected by chance (overdispersed, i.e. evenly spread across the phylogenetic tree), whereas communities structured by habitat filtering would be more closely related (i.e. clustered). Finally, phylogenetic β-diversity patterns were explored (changes in species composition along environmental/spatial gradients and among communities), as well as the qualitative salinity effect by comparing thalassohaline (i.e. salt composition similar to seawater) and athalassohaline (salt composition different from seawater) water masses. Inland (i.e. isolated water bodies understood as 'pond as a patch') and marine sites (well-connected environments) offered a useful contrasting framework to explore the most plausible mechanisms responsible for the phylogenetic diversity, structure and β-diversity patterns observed for the different bacterial groups.

4.2 Methods

4.2.1 Data set characteristics

After a bibliographic search we found 34 different locations that carried out extensive clone libraries (> 40 valid sequences of the bacterial 16S rRNA gene per site) of surface or epilimnetic waters, for a total of 7,154 sequences. However, some data sets were already available in databases as non-redundant 97% identity-grouped sequences, a consensus value for delimiting bacterial species, but not without concerns (Rosselló-Mora & Amann, 2001), and only 1 sequence for each 97% cut-off group was reported (see Appendix). Therefore, to minimize the bias introduced by the different sampling ef-

forts for each clone library and to agree on a conservative phylogenetic criterion, we grouped the remaining data at 97% identity using the percentage sequence identity (PSI) algorithm of the online program FastGroupII (http://biome.sdsu.edu/fastgroup/). Sequences shorter than 300 nucleotides were also discarded. Therefore, a non-redundant, consistent (all sequences obtained from the same cloning methodology), and balanced (equal number of marine sites and lakes, although with different sampling effort in each case that we corrected using randomized subsamples, see below) data set of 4,495 sequences was retained for further analysis (see Appendix for more detailed information).

4.2.2 16S rRNA gene sequence data analysis

The 16S rRNA gene pool was automatically aligned with the NAST alignment tool (DeSantis et al., 2006a). Next, the aligned data set was imported into the ARB software package (Ludwig et al., 2004) and was added by parsimony to the optimized Greengenes tree provided by default (>130,000 sequences, May 2007 release). Each 16S rRNA gene sequence was assigned to a bacterial phylogenetic group (e.g. *Alpha-*, *Beta-*, *Gamma-*, and *Deltaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, *Bacteroidetes*, *Firmicutes*, *Planctomycetes*, and *Verrucomicrobia*) in a hierarchy based on the Ribosomal Database Project (http://rdp.cme.msu.edu).

4.2.3 Community phylogenetic analyses

PD for each community was calculated as the sum of the branch lengths associated with the overall 16S rRNA sequences obtained from that sample. Thus, PD takes into account the number of sequences (i.e. species richness) and the phylogenetic relationship among them (Faith, 1992). Richness and Shannon diversity index were calculated for each site at the taxonomic level selected. To correct for unequal sample sizes, we calculated the mean of the richness, the Shannon diversity, and PD of 1,000 randomized subsamples for each community (Bryant et al., 2008). The subsample size was the number of sequences present in the smallest community.

Subsequent analyses were carried out selecting the most abundant groups (to make the phylogenetic inference reliable) present at all sites (to give consistency to the global patterns observed). Members of the *Bacteroidetes* were not separated in smaller phylogenetic units because *Sphingobacteria* and *Flavobacteria* were mostly absent from marine and inland waters, respectively. To assess driving processes in bacterioplankton community assemblages, we compared the different tree topologies using two indices, i.e. the NRI and the

NTI (Webb, 2000). These indices measured the degree of phylogenetic relatedness of the different taxa from a sample relative to the pool of taxa from all the samples. Relatedness information provides a different window into bacterial communities than does information concerning richness or taxonomic composition. NRI reveals wide patterns across the phylogeny, while NTI focuses on terminal taxa. High and positive values of these indices indicate clustering of taxa across the overall phylogeny, whereas low or negative values indicate overdispersion of taxa across the phylogeny. In other words, a positive value indicates that bacteria tended to co-occur with other bacteria that were more closely related than expected by chance (for more details see Webb, 2000; Horner-Devine & Bohannan, 2006). We calculated NRI and NTI using Phylocom 3.41 (www.phylodiversity.net/phylocom).

Distance matrices among environments were calculated using the UniFrac metric (Lozupone & Knight, 2005). The UniFrac distance is calculated as the percent of branch length leading to descendants from only one of the locations represented in the phylogenetic tree (Lozupone & Knight, 2007). Hence, UniFrac is a β-diversity metric that quantifies community similarity based on phylogenetic relatedness.

4.2.4 Statistical analyses

We used standard and partial Mantel tests to determine the correlation between similarity in microbial communities (based on UniFrac matrices) and geographic (S) and environmental (E) matrices. Genetic databases usually lack environmental metadata associated with sequence information, and this limitation should be corrected in the future. Therefore, we tried to compile environmental data from the original publications and by requesting information from the authors. Unfortunately, it was not possible to obtain sufficient environmental variables simultaneously for all sites to more finely contextualize the information. Thus, we selected general variables traditionally used in limnology and oceanography. Lakes are usually classified by their geomorphology, as it encapsulates most of their variation and correlates with other fundamental variables. For marine sites, we used the physico-chemical parameters that define broad water masses. Data used are shown in Appendix. The spatial matrix (S) was calculated following detailed geographic coordinates and the Earth's curvature. The environmental matrix (E) was constructed with general geomorphological features for lakes (i.e. area, depth, elevation), whereas for oceanic waters we used physico-chemical data (i.e. temperature, salinity, dissolved oxygen) provided by the authors or estimated from the NOAA World Ocean Atlas (www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html). We ran non-metric

multidimensional scaling (NMDS) to represent the ordering relationships obtained from the UniFrac distance matrices. To further explore environmental effects, each sample was classified into two different categories: one based on salt quality or composition (i.e. thalassohaline and athalassohaline) and the other on salt quantity or concentration (i.e. freshwater, saline, hypersaline). Thalassohaline habitats have salt composition similar to seawater, whereas athalassohaline environments are inland salt ponds with salt composition different from seawater. To assess the significance of the salinity categories, we performed analyses of similarities (ANOSIM) based on 1,000 permutations. The ANOSIM R statistic is based on the difference of mean ranks between groups and within groups and ranges from 0 (no separation) to 1 (complete separation) (Clarke, 1993). Statistical analyses were carried out in the R environment (www.r-project.org) using the vegan package (Oksanen et al., 2009).

4.3 Results

The number of 16S rRNA gene sequences analyzed for each site, as well as geographical coordinates, available environmental description, and reference in the literature for each of the aquatic environments surveyed in the present study are presented in the Appendix. Overall, 18 inland water bodies of different sizes, altitudes, and trophic status and 16 mainly coastal marine sites were analyzed using ca. 4,500 gene sequences. Clone libraries indicated that surface bacterioplankton was mainly dominated by 2 phyla, i.e. *Proteobacteria* (especially *Alpha-*, *Beta-*, and *Gammaproteobacteria*) and *Bacteroidetes* (especially *Flavobacteria* and *Sphingobacteria*). Inland waters harbored more bacterial groups (p-value = 0.017, t-test), and these were more diverse (p-value < 0.001, t-test) than in marine waters. Additionally, lakes presented greater (but not significant) PD than marine samples (p-value = 0.079, t-test). Figure 4.1 shows the respective fractions of 16S rRNA gene sequence types in freshwater and marine habitats, testing for significant (i.e. *Alpha-*, *Beta-*, and *Gammaproteobacteria*, *Actinobacteria*, *Firmicutes*, *Flavobacteria*, and *Sphingobacteria*; p-value < 0.05, t-test) and not significant (e.g. *Cyanobacteria*, *Deltaproteobacteria*, *Planctomycetes*, *Verrumicrobia*) differences between the pairs of frequencies. Glöckner et al. (1999) used whole-cell fluorescence in situ hybridization (FISH) to indicate for the first time significant differences in the composition of marine and lake bacterioplankton with a limited number of samples and environments. Here, we used a larger data set obtained by cloning and sequencing that confirmed more detailed striking differences in global relative abundance and distribution for the most abundant groups in the two types of environments. In

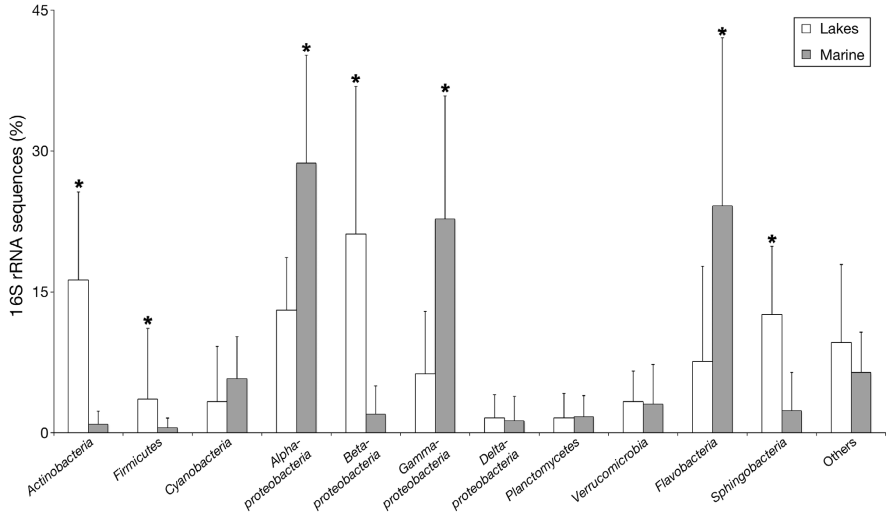


Figure 4.1: Relative abundance of bacterial groups present in the different lakes and marine habitats examined using the 4,495 16S rRNA gene sequence data set (grouped at 97% identity). Error bars indicate one standard deviation. Groups with significant differences between pairs of frequencies (p-value < 0.05, t-test) are labeled with an asterisk.

inland waters, *Betaproteobacteria*, *Actinobacteria*, *Sphingobacteria*, and *Firmicutes* were the most abundant, while in marine bacterioplankton *Alphaproteobacteria*, *Flavobacteria*, and *Gammaproteobacteria* dominated (Figure 4.1).

For the analysis of processes driving community phylogenetic structure, both NRI and NTI indicated that bacterial assemblages had significantly higher phylogenetic clustering than expected by chance (Table 4.1), suggesting that communities were structured by habitat filtering. Although there was some degree of variation, such clustered structure was consistent for the different environments and different bacterial groups considered. Marine habitats showed a higher percentage of clustered sites (approximately 55% for NRI and 77% for NTI) than lakes (approximately 33% for NRI and 57% for NTI) (p-value = 0.031 for NRI and p-value = 0.079 for NTI, t-test).

At a global scale and pooling together the bacterial groups as a whole (Figure 4.2A), the ordination analysis of the phylogenetic β-diversity matrix grouped the aquatic environments into 3 salinity concentration categories (R = 0.92, p-value < 0.001, ANOSIM). Salt quality or salinity composition was also important (R = 0.87, p-value < 0.001), although hypersaline environments tended to cluster together with either thalassohaline and athalassohaline environments. We observed the same pattern for *Bacteroidetes* (R = 0.92, p-value

Table 4.1: Summary of net relatedness index (NRI) and nearest taxa index (NTI) results. Numbers are the percentages of significant structured sites found for each group. Significance was assessed at values of 0.05 and 0.10, and was tested with a null model (999 random permutations across the entire phylogeny). A two-tailed test evaluated the rank of observed values.

Community	NRI		NTI	
	Overdispersed	Clustered	Overdispersed	Clustered
All aquatic environments				
All bacteria	3	53	0	91
Bacteroidetes	3	33	0	67
Alphaproteobacteria	3	31	3	59
Gammaproteobacteria	9	59	0	73
Lakes				
All bacteria	0	44	0	89
Bacteroidetes	0	6	0	53
Alphaproteobacteria	6	19	6	38
Gammaproteobacteria	0	50	0	50
Actinobacteria	7	33	0	47
Betaproteobacteria	0	43	7	64
Marine habitats				
All bacteria	6	63	0	94
Bacteroidetes	6	63	0	81
Alphaproteobacteria	0	44	0	81
Gammaproteobacteria	13	63	0	81
Cyanobacteria	0	42	0	50

< 0.001 for salt concentration, and R = 0.80, p-value < 0.001 for salt composition) (Figure 4.2B). However, β-diversity patterns for *Alphaproteobacteria* and *Gammaproteobacteria* (Figure 4.2C, D) appeared less sensitive to salt concentration (R = 0.81, p-value < 0.001 and R = 0.73, p-value < 0.001, respectively) than to salinity composition (R = 0.91, p-value < 0.001 and R = 0.86, p-value < 0.001, respectively).

Overall, when all aquatic environments were assembled and their phylogenetic β-diversity contextualized, the correlation analysis showed the significance of the environmental component (after fixing the spatial component by partial Mantel tests) for the different groups considered (Table 4.2). After splitting environments into lakes and marine habitats, environmental determinism was still the best explanatory dynamic for most of the groups, but some relevant exceptions appeared for inland bacterioplankton. Inland *Actinobacteria* and *Betaproteobacteria* showed, in addition to the environmental component, a significant contribution from the spatial component (after fixing the environmental component by partial Mantel tests) (Table 4.2). The geographical distribution merely explained the correlations found in inland *Gammaproteobacteria*, but not in the marine counterparts (Table 4.2). In Table 4.2 we also tentatively assigned the most appropriate metacommunity model following Cot-

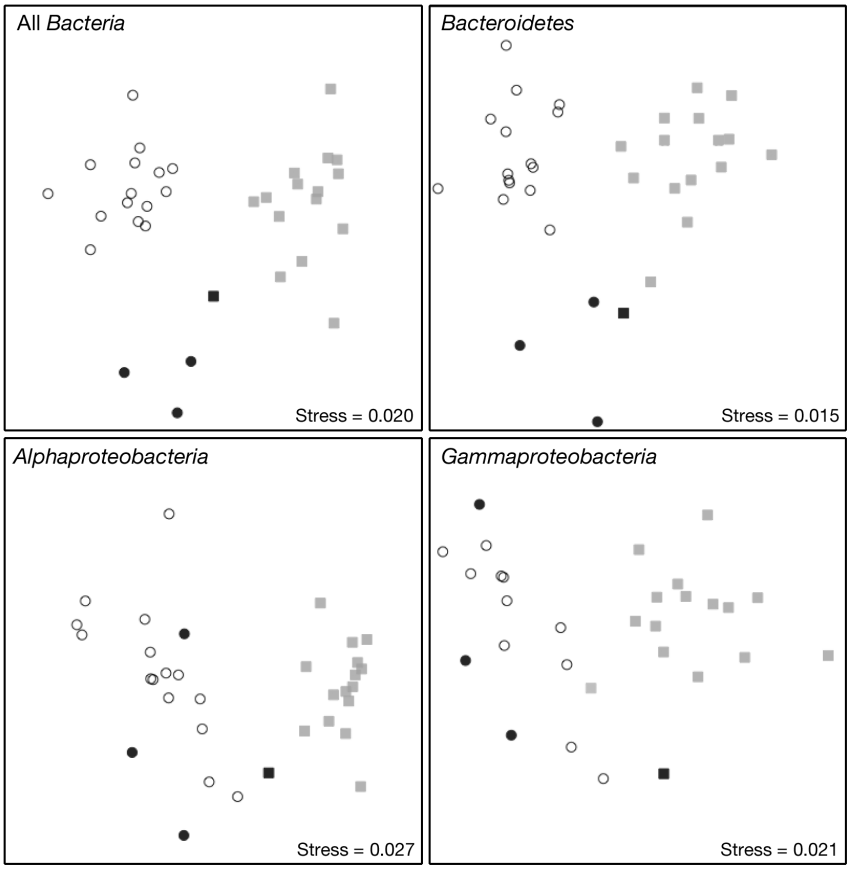


Figure 4.2: Non-metric multidimensional scaling (NMDS) plots based on the UniFrac matrices for all *Bacteria*, *Bacteroidetes*, *Alphaproteobacteria* and *Gammaproteobacteria* (i.e. those groups present at all sites with enough available sequences). Circles: inland salt composition (athalassohaline); squares: marine salt composition (thalassohaline); open symbols: freshwater; grey: saline water; black: hypersaline water. Normalized stress values are shown for each plot. Stress is a measure of the mismatch between original distance measures (the UniFrac distance) and distance in ordination space.

tenie (2005), bearing in mind that the assignment of different metacommunity models is dependent on the p-value used to define statistical significance and would lead to non-identical β-diversity patterns (in Table 4.2), see bold and italic values for significant [p-value < 0.05] positive correlations and boldfaced print for the more relaxed [p-value < 0.1] positive correlations that we used for discerning patterns). Both all aquatic environments and marine habitats agreed with the SS (species-sorting) model in all cases. Freshwater *Actinobacteria* and *Betaproteobacteria* showed, in addition, a signal consistent with the ME metacommunity model. Inland *Gammaproteobacteria*, in turn, formed a spatial pattern equally explained by neutral or patch dynamics (see more details in the "Discussion" section).

4.4 Discussion

The correlation approach used in the present study showed community assembly based on environmental selection as a relevant potential mechanism for bacterioplankton biogeographical structure at a global scale. Recently, several studies have emphasized the importance of environment in shaping aquatic bacterial communities at the local scale (Langenheder & Ragnarsson, 2007; Van der Gucht et al., 2007; Sommaruga & Casamayor, 2009). Hence, environmental adaptations explain the abundance and distribution of bacterial assemblages at multiple scales of study.

Community assembly rules are intriguing for microbial ecologists, due to the lack of knowledge about the dispersal potential and effective niche space for bacteria. Some recent studies suggest that dispersal ability is probably not high for all bacteria (Papke & Ward, 2004; Hervàs et al., 2009). Here, we consistently observed distinct bacterial composition in the plankton of marine and inland habitats even using broad taxonomic classification. These results match previous findings by whole-cell FISH (Glöckner et al., 1999), where a limited number of samples (4 lakes, 4 marine sites) were surveyed using group-specific probes (*Alpha-*, *Beta-* and *Gammaproteobacteria*, *Bacteroidetes*, and *Planctomycetes*). Essentially, the same results were obtained here with a larger data set and with a different approach (DNA extraction, PCR amplification, and gene cloning and sequencing). Thus, although pure quantitative composition cannot be inferred from clone libraries, these molecular techniques provide semi-quantitative information that can be used to assess such general trends as those observed in our meta-analytical study. In particular, we consistently observed that *Actinobacteria* and *Sphingobacteria*, in addition to the previously known *Betaproteobacteria*, were mostly absent from marine habitats and hypersaline lakes, while they were abundant in freshwater lakes, indicat-

Table 4.2: Summary of the results obtained with the UniFrac matrices against the environmental ([E]) and the spatial ([S]) components, and tentative assignment to the most appropiate metacommunity model following Cottenie (2005). Mantel correlations (r_M) and p-values obtained after 999 permutations. Bold and italic values: significant (p-value < 0.05) positive correlations; bold: partially significant (p-value < 0.1) positive correlations; not highlighted values: non-significant or negative correlations. [E] and [S]: standard Mantel tests; [E|S], [S|E]: partial Mantel tests with environment independent of space and space independent of environment, respectively; SS: species-sorting; NM: neutral model; PD: patch dynamics; ME: mass-effects.

Community	[E]		[S]		[E | S]		[S | E]		Metacommunity type
	r_M	p-value	r_M	p-value	r_M	p-value	r_M	p-value	
All aquatic environments									
All bacteria	***0.521***	***0.001***	0.040	0.241	***0.520***	***0.001***	-0.011	0.469	SS
Bacteroidetes	***0.475***	***0.001***	0.080	0.119	***0.471***	***0.001***	0.039	0.275	SS
Alphapoteobacteria	***0.493***	***0.001***	-0.047	0.206	***0.501***	***0.001***	-0.108	0.022	SS
Gammaproteobacteria	***0.563***	***0.001***	**0.108**	**0.063**	***0.558***	***0.001***	0.062	0.187	SS
Lakes									
All bacteria	***0.237***	***0.004***	0.157	0.153	***0.203***	***0.006***	0.095	0.249	SS
Bacteroidetes	***0.192***	***0.016***	0.051	0.310	***0.186***	***0.014***	-0.005	0.546	SS
Alphaproteobacteria	**0.110**	**0.090**	-0.087	0.270	***0.141***	***0.039***	-0.125	0.171	SS
Gammaproteobacteria	***0.228***	***0.018***	***0.576***	***0.001***	0.068	0.253	***0.546***	***0.002***	NM/PD
Actinobacteria	***0.168***	***0.047***	**0.207**	**0.071**	**0.155**	**0.062**	**0.197**	**0.083**	SS+ME
Betaproteobacteria	***0.253***	***0.007***	***0.326***	***0.026***	***0.234***	***0.008***	***0.312***	***0.034***	SS+ME
Marine habitats									
All bacteria	***0.388***	***0.012***	-0.236	0.033	**0.364**	**0.067**	-0.237	0.900	SS
Bacteroidetes	***0.500***	***0.009***	-0.109	0.228	***0.486***	***0.041***	-0.125	0.721	SS
Alphaproteobacteria	***0.430***	***0.010***	-0.239	0.030	**0.411**	**0.057**	-0.187	0.857	SS
Gammaproteobacteria	***0.392***	***0.007***	-0.192	0.073	**0.366**	**0.056**	-0.319	0.982	SS
Cyanobacteria	**0.374**	**0.089**	-0.076	0.294	0.358	0.111	-0.081	0.628	Not found

ing that salt concentration may drive their potential distribution. Although a large number of environmental parameters covary with salinity, consistent differences have been shown in microbial assemblages along the quantitative salinity gradient from seawater to NaCl saturation (Casamayor et al., 2000a) that were higher than along qualitative gradients (i.e. thalassohaline vs. athalassohaline sites; Demergasso et al., 2004). In addition to the previously reported fact that salinity was the major filtering factor for bacterial and archaeal community assemblages (Lozupone & Knight, 2007; Auguet et al., 2010), our analysis also detected different responses to salt composition among the dominant bacterioplankton groups. We observed that, while β-diversity patterns for *Bacteroidetes* were mostly shaped by salinity concentration, patterns among *Alphaproteobacteria* and *Gammaproteobacteria* were controlled by salt composition.

Besides the difference in community composition between inland and marine environments, contrasting geographical β-diversity patterns were observed. Such disparity could be related to the connectivity among habitats in the marine continuum versus lakes as islands within a sea of land (Reche et al., 2005, and references therein). Accordingly, marine habitats showed β-diversity patterns not explained by the spatial component. This is in agreement with the concept of the ocean surface as a contiguous and relatively homogeneous environment for bacteria that facilitates dispersion, and may lead to the lack of geographical patterns in marine bacteria (Pommier et al., 2007). Nonetheless, we found that local environmental factors (not related to spatial distance) were still very powerful constraints in structuring marine bacterioplankton assemblages. In contrast, inland water bodies had a patchy distribution more likely to exhibit dispersal constraints in addition to environmental factors. Other studies have found a significant spatial component operating at the regional scale for lakes and ponds (e.g. Reche et al., 2005, 2007; Langenheder & Ragnarsson, 2007). The disconnected and relatively heterogeneous nature of lakes (Papke & Ward, 2004) may have promoted the higher bacterial richness and diversity observed in our study. Recently, freshwater environments have also been unveiled as one of the largest reservoirs of archaeal diversity (Auguet et al., 2010). Hence, inland water bodies appear as promising environments for biological novelty in terms of phylogenetic diversity and metabolic potential (e.g. Demergasso et al., 2008).

The analysis of the phylogenetic β-diversity for all aquatic environments together and marine habitats agreed with the SS model for all the groups examined (Table 4.2), as interpreted by Cottenie (2005) emphasizing environmental adaptations to explain the distribution of species. For lakes, in addition to an environmental signal, a relationship between community similarity and distance was detected for inland *Actinobacteria* and *Betaproteobacte-*

ria, consistent with the ME metacommunity model, which mainly focuses on the effect of high immigration on local dynamics, but also may reflect dispersal restrictions (or the influence of unmeasured environmental parameters; Table 4.2). In fact, SS + ME dynamics may indicate a metacommunity with species sorting and either high dispersal (which are true mass effects) or low dispersal between at least some local communities (species sorting and dispersal limitation), although it is difficult to find out which of those is the case (Ng et al., 2009). MEs emphasize the role of spatial dynamics on community structure in a system where species can be rescued from local competitive exclusion in communities where they are poor competitors by immigration. This will result in spatial patterns occurring independently of environmental patterns (Cottenie, 2005). For this study and the spatial scales that are considered, species sorting coupled with dispersal limitation is the more likely scenario. True mass effects would mean that dispersal rates between the different sites are higher than the internal growth rates of local populations, and this is very unlikely given that most locations are not part of the same catchment area. However, both *Betaproteobacteria* and *Actinobacteria* tend to co-occur and are very abundant in nearly all lakes, suggesting continuous feeding with allochthonous bacteria from the catchment area, but atmospheric dust depositions should also be considered in remote freshwater ecosystems where these two groups have been detected with the potential to grow (Hervàs et al., 2009). *Gammaproteobacteria*, in turn, formed a spatial pattern equally explained by neutral or patch dynamics. Many members of this phylogenetic group are typical copiotrophs, adapted to relatively high nutrient concentrations such as those in culturing media, phytoplankton blooms, or coastal areas (e.g. Agogue et al., 2005; Fuchs et al., 2007) that show sporadic peaks of abundance along river–estuary transects (e.g. Bouvier & del Giorgio, 2002) or in lakes (e.g. Casamayor et al., 2002). Thus, if they have a feast and famine strategy and frequently peak in abundance in response to nutrient pulses, this would be rather an indication of species sorting than neutral or patch dynamics. However, some of the occurrences of suitable conditions may have an unpredictable and sporadic nature, and the pattern observed might be related to such local sporadic events that produce suitable conditions to proliferate or, alternatively, to integrate other non-considered variables in the analysis, such as the temporal variation hidden in the spatial signal (see below).

A major focus in the field of biogeography is the relative influence of current environmental versus historical factors on present day distribution patterns. In the case of microorganisms, large population size and high dispersal may blur the potential effect of historical legacy on present community patterns (Fuhrman et al., 2008). For that reason, community phylogenetic

structure is more likely to be generated by ecological rather than historical processes (i.e. diversification within a region causing many taxa to be, on average, more related to each other rather than taxa outside the region; Webb et al., 2002). Presuming phylogenetic niche conservatism (i.e. related taxa are more ecologically similar), phylogenetic clustering indicates environmental selection, while overdispersion can be considered the result of competition between closely related species or facilitation between distantly related taxa. However, phylogenetic resolution and sampling scale may influence the observed patterns (for a discussion see Horner-Devine & Bohannan, 2006). In the present study, we found that the closest bacterioplankton relatives tend to be in the same habitat (habitat filtering, owing to the sharing of adaptations needed for a particular habitat) in agreement with recently published reports for bacterial communities (Horner-Devine & Bohannan, 2006; Newton et al., 2007; Bryant et al., 2008). The broad scales of environmental gradients, geography, and taxonomy explored in our study may have concealed overdispersed patterns (Horner-Devine & Bohannan, 2006, and references therein). Despite this limitation, analytical exploration on such broad scales is still scarce for microorganisms (but see Fierer & Jackson, 2006; Pommier et al., 2007) and relevant information is still to be unveiled. Some other methodological limitations inherent to the shortcomings in databases are also obvious in the work presented here, such as the rationale for the environmental variables chosen or the strong sampling bias toward northern temperate regions. Although we tried to homogenize the data sets by selecting clone library studies with high numbers of 16S rRNA gene sequences per site (> 40) and by selecting superficial waters, other factors could have escaped our analysis, such as the temporal component that may lead to some predictable patterns (Fuhrman et al., 2006) or the influence of biotic factors on bacterial community composition (Kent et al., 2007). Other problems are intrinsic to the molecular ecology approach; for example, (1) PCR cloning focuses mostly on abundant taxa and may cause bias against rare taxa, resulting in semi-quantitative information, or (2) dereplicating 16S rRNA libraries at the species level (97% identity) implicitly loses the microdiversity inherent in most bacterial lineages. A minor sequence change in such a highly conserved gene may conceal relevant differences in whole genome content and, thus, in ecological capability, and there are some cases of different bacterial species with identical 16S rRNA genes (see a review in Rosselló-Mora & Amann, 2001). In the worst case, methodological limitations will not affect the general patterns observed here because we focused on conservative criteria, and we dealt with the most abundant bacterial groups detected. The abundant members cover the part of the microbial biodiversity spectrum that includes the active units of the community (so-called "core species") that drive most ecosystem functions, whereas those below nu-

merical detection thresholds of molecular techniques form a seed bank of rare species that grow slowly or not at all. The rare species on the whole cannot be completely unveiled with traditional PCR methods, although biogeographical approaches would be greatly enhanced by increasing the genetic information deposited in databases and by improving the coverage of sequencing techniques (Galand et al., 2009, 2010). In addition, there is ample room for community phylogenetic approaches involving more specific bacterial clades and narrower scales (e.g. Newton et al., 2007).

Only very recently has it been observed that microorganisms may have similar ecological patterns to their macroorganism counterparts in terms of the taxa–area relationship (Horner-Devine et al., 2004; Reche et al., 2005, 2007), the latitudinal and elevational gradients of richness (Pommier et al., 2007; Bryant et al., 2008; Fuhrman et al., 2008), or the community assembly rules (Horner-Devine et al., 2007). One of the most puzzling aspects of ecology is the wide range of spatial, temporal, and taxonomical scales at which patterns may arise. As a result of their interrelationship, some patterns emerge as processes on a lower scale level (Levin, 1992). Moreover, it has been shown that microbial groups have ecological idiosyncrasies and habitat preferences that are stable over time (Von Mering et al., 2007). Therefore, microbial biogeography, although traditionally disregarded, is a particularly motivating field of study, due to the wide range of scales involved (Dolan, 2006).

Our empirical meta-analysis, which used a hierarchical phylogenetic approach, reinforces the niche explanation for the global distribution of surface bacterioplankton assemblages. The results indicate differences among phylogenetic groups that deserve further investigations, and the patterns we observed might help in the future as a basis upon which to design studies and experiments to test various aspects related to community assembly theory. With current molecular techniques, we cannot definitely state whether everything is everywhere or not (no dispersal and colonization barriers for any bacterial cell), but certainly there is a strong aspect of environmental determinism shaping aquatic microbial assemblages on a global scale, and environmental forces are acting singularly on different bacterial groups.

Acknowledgements

We are thankful to authors who provided valuable data for the analysis, to Micah Hamady and Campbell O. Webb for technical assistance with UniFrac and Phylocom, respectively, to Antonio Fernàndez-Guerra for computing assistance, and to anonymous reviewers for helpful comments and improvements of earlier versions of this manuscript. This research was sup-

ported by Grant AERBAC 079/2007 to EOC from the Spanish Ministerio de Medio Ambiente (MARM), and CONSOLIDER-INGENIO 2010 Grant GRACCIE CSD2007-00004 from the Spanish Ministerio de Ciencia e Innovación (MICINN). AB was supported by a FPU-MICINN predoctoral scholarship.

Appendix [2]

[2]See more Supplementary Information in Barberán & Casamayor (2010).

Table 4.3: Number of 16S rRNA gene sequences (clustered at 97% identity), coordinates, environmental description and references of the lakes analyzed in this study.

Site	Number of sequences	Coordinates	Area (Ha)	Depth (m)	Elevation (m)	Description	Reference
Blankaart	46	50°59′N,02°51′E	32	1	1	Peat, shallow	Van der Gucht et al. (2005)
Maten12	61	50°57′N,05°27′E	3.2	0.7	44	Peat, shallow	Van der Gucht et al. (2005)
Maten13	81	50°57′N,05°27′E	3.3	0.4	44	Peat, shallow	Van der Gucht et al. (2005)
Visvijver	47	50°59′N,02°52′E	0.6	1	1	Peat, shallow	Van der Gucht et al. (2005)
Ekoln	72	59°45′N,17°36′E	2000	40	3	Glacial	Eiler & Bertilsson (2004)
Erken	73	59°25′N,18°15′E	2370	21	1	Glacial	Eiler & Bertilsson (2004)
Limmaren	91	59°44′N,18°44′E	650	8	5	Glacial	Eiler & Bertilsson (2004)
Vallentunasjon	56	59°29′N,18°02′E	620	5	10	Glacial	Eiler & Bertilsson (2004)
Crater	57	42°56′N,122°06′W	5320	589	1882	Volcanic	Urbach et al. (2001)
Fuchskuhle	58	53°10′N,13°02′E	2	6	59	Peat, shallow	Glöckner et al. (2000)
Fazda	54	30°19′N,30°24′E	113	3	1	Hypersaline, shallow	Mesbah et al. (2007)
Hamra	65	30°23′N,30°19′E	63	6	2	Hypersaline, shallow	Mesbah et al. (2007)
Crystal	103	46°00′N,89°36′W	0.56	3	500	Peat, shallow	Newton et al. (2006)
Tebenquiche	50	23°08′S,68°15′W	125	2	2350	Hypersaline, shallow	Demergasso et al. (2008)
Heywood	78	60°43′S,45°38′W	4.17	6	4	Glacial	Pearce et al. (2005)
Taihu	190	31°24′N,120°13′E	233800	4	3	Riverine, shallow	Wu et al. (2007)
Dongting	45	29°19′N,112°57′E	28200	31	28	Riverine	Sekiguchi et al. (2002)
Poyang	53	29°00′N,115°30′E	358500	25	12	Riverine	Sekiguchi et al. (2002)

Table 4.4: Number of 16S rRNA gene sequences (clustered at 97% identity), coordinates, environmental description and references of the marine habitats analyzed in this study.

Site	Number of sequences	Coordinates	Temperature (ºC)	Salinity (psu)	Dissolved O_2 ($ml \cdot l^{-1}$)	Description	Reference
Cape Town	194	34º15′S,17º53′E	18.06	35.44	5.58	Coastal	Pommier et al. (2007)
San Diego	220	32º53′N,117º23′W	16.98	33.5	5.87	Coastal	Pommier et al. (2007)
Hawaii	243	21º10′N,157º55′W	25.15	34.94	4.81	Coastal	Pommier et al. (2007)
Baffin Bay	90	69º15′N,53º33′W	4	32.69	8.32	Coastal	Pommier et al. (2007)
Fiji	364	18º10′S,178º12′E	26.16	35.12	4.57	Coastal	Pommier et al. (2007)
Sargasso	298	32º13′N,64º39′W	23	36.49	4.97	Coastal	Pommier et al. (2007)
Chile	364	36º29′S,73º10′W	14.04	33.73	5.81	Coastal	Pommier et al. (2007)
Arctic	168	78º00′N,02º02′W	0.21	33.73	8.2	Open ocean	Pommier et al. (2007)
Sydney	176	34º07′S,151º12′E	20.79	35.6	5.04	Coastal	Pommier et al. (2007)
Blanes	115	41º40′N,02º48′E	15.29	37.82	5.51	Coastal, Mediterranean	Alonso-Sáez et al. (2007)
North Sea	55	55º08′N,01º16′W	10.13	34.4	5.98	Coastal	Franklin et al. (2005)
Long Island	71	40º41′N,73º05′W	12.16	31.34	6.32	Coastal	Kelly & Chistoserdov (2001)
Ionian	256	36º30′N,15º50′E	19.88	38.05	5.25	Coastal, Mediterranean	Zaballos et al. (2006)
Plum Island	404	42º49′N,70º47′W	16	31.69	6.93	Coastal	Acinas et al. (2004)
Plymouth	147	50º15′N,04º13′W	12.35	35.06	6.27	Coastal	O'Sullivan et al. (2004)
Santa Pola	50	38º12′N,0º36′W	25	200	5.03	Coastal, hypersaline, Mediterranean	Benlloch et al. (2002)

5

Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas

Resumen

Los dominios *Bacteria* y *Archaea* representan la vasta mayoría de la biodiversidad. No obstante, la interacción entre los procesos ecológicos y evolutivos en el mundo microbiano permanece desconocida. En este estudio, se han explorado los patrones de comunidades planctónicas de bacterias que habitan lagos estratificados con capas óxicas/anóxicas y euxinia. Se examinó si esta estratificación vertical es promotora del endemismo en capas profundas mediante el análisis de secuencias del gen ribosomal del 16S. La similaridad en la composición mostró que las comunidades de la misma capa de agua eran más parecidas entre lagos que las comunidades de diferentes capas del mismo lago. Además, el hipolimnion anóxico presentó mayor β-diversidad que el epilimnion óxico. Una mayor β-diversidad puede ser atribuible a una baja dispersión y escasa conectividad entre territorios. Paralelamente, mientras que las aguas superficiales exhibieron un componente espacial significativo, en el caso de las capas profundas, el componente significativo fue el ambiental. Por lo tanto, diferentes mecanismos ecológicos actúan simultáneamente en la misma masa de agua. En general, el endemismo en bacterias es probablemente más común que lo supuesto con anterioridad, particularmente en hábitats de aguas dulces aislados y ambientalmente heterogéneos.

Abstract [1]

Bacteria and *Archaea* represent the vast majority of biodiversity on Earth. The ways that dynamic ecological and evolutionary processes interact in the microbial world are, however, poorly known. Here, we have explored community patterns of planktonic freshwater bacteria inhabiting stratified lakes with oxic/anoxic interfaces and euxinic (anoxic and sulfurous) water masses. The interface separates a well-oxygenated upper water mass (epilimnion) from a lower anoxic water compartment (hypolimnion). We assessed whether or not the vertical zonation of lakes promoted endemism in deeper layers by analyzing bacterial 16S rRNA gene sequences from the water column of worldwide distributed stratified lakes and applying a community ecology approach. Community similarity based on the phylogenetic relatedness showed that bacterial assemblages from the same water layer were more similar across lakes than to communities from different layer within lakes and that anoxic hypolimnia presented greater β-diversity than oxic epilimnia. Higher β-diversity values are attributable to low dispersal and small connectivity between community patches. In addition, surface waters had significant spatial but non-significant environmental components controlling phylogenetic β-diversity patterns, respectively. Conversely, the bottom layers were significantly correlated with environment but not with geographic distance. Thus, we observed different ecological mechanisms simultaneously acting on the same water body. Overall, bacterial endemicity is probably more common than previously thought, particularly in isolated and environmentally heterogeneous freshwater habitats. We argue for a microbial diversity conservation perspective still lacking in the global and local biodiversity preservation policies.

5.1 Introduction

Multicellular animals and plants may face geographic barriers to migration and dispersal, enabling isolated populations to diverge into different species (allopatric speciation). Microorganisms, in turn, can be easily transported to very distant places (e.g. Hervàs et al., 2009), and total dispersal success has been postulated based on their small size, vast population sizes, long survival, low extinction rate, and astronomic global numbers that may lead to microbial ubiquity (Finlay, 2002). As ubiquity limits rates of local speciation and extinction, a relative homogeneity in the composition of microbial species is expected worldwide. The importance of geographical and environmental bar-

[1] See original publication in Barberán & Casamayor (2011).

riers to divergence in microbial populations is, however, a subject of great debate, and whereas some species show evidence of allopatric divergence, some others do not (see a recent review in Whitaker, 2006). For bacteria, some phylotypes show, in fact, evidences of a worldwide distribution (e.g. Barberán & Casamayor, 2010), but some bacterial genera have different species in the Arctic than in the Antarctic sea ice or geographical barriers have been shown for populations of hyperthermophilic archaea (Whitaker et al., 2003), suggesting that dispersal is simply prevented by the geographical separation of populations, and endemisms do exist among bacteria and archaea. Certainly, we have only a vague idea on what are the evolutionary parameters that lead to barriers in some microbial species, but not in others (Whitaker, 2006).

Our ignorance about which environments are more likely to hold a large microbial diversity and which ecosystems may be islands for microorganisms is also large. We have to strain our imagination in order to think what a barrier can be for a bacterium and on the mechanisms that drive divergence among microbial lineages. If endemic microbial species are common, the total number of species should be large, whereas for a cosmopolitan distribution, we may expect low number of bacterial species. Unfortunately, present estimations on the number of bacterial species range several orders of magnitude showing a large degree of uncertainty being far from a consensus value (Pedrós-Alió, 2006). We also have problems in disentangling mechanisms that result in bacterial biogeographic patterns driven by environmental selection from a panmictic (fully mixed) pool to those driven by local evolution in geographically isolated populations (allopatry). Physical isolation strongly influences speciation and both speciation and extinction strongly influence species richness. Understanding the evolutionary ecology of microorganisms has also practical relevance for designing conservation strategies to protect essential habitats with endemic microbial populations (Souza et al., 2006), a concern ignored in management actions and policies aimed at preserving biodiversity and making use of the resources and services it provides. However, since most microbiologists remain ambiguous on the definition of a bacterial species, investigating the processes of speciation in environmental communities is certainly a difficult goal.

Here, we have explored community patterns of planktonic freshwater bacteria inhabiting stratified lakes with oxic/anoxic interfaces along the water column. The interface separates a well-oxygenated upper water mass (epilimnion) from a lower anoxic water compartment (hypolimnion). Within the large heterogeneity of inland waters, stratified lakes appear particularly suited to unravel how environmental heterogeneity and dispersal influence bacterial community composition in a large spatial context and the concordance of those patterns between water compartments. The variety of metabolisms

and microbial communities occurring in separated density layers of stratified lakes is large (Casamayor et al., 2001, 2000b; Konopka et al., 1999; Llirós et al., 2008) and anaerobic microbes that live in the deep anoxic water compartments are separated from the atmosphere by oxic water layers and have a less marked temporal variation than their epilimnetic counterparts (Casamayor et al., 2002). Thus, perhaps, anoxic layers of lakes may act as islands in an aerobic world. This would reduce dispersion of anaerobic bacteria from one lake to another lake and would favor the presence of endemic species. Additionally, selection may also be imposed by differences in environmental conditions among hypolimnia. In the present meta-analysis, we have explored the patterns of bacterioplankton assemblages by the analysis of the phylogenetic community structure (16S rRNA gene sequences) under a macroecological perspective. We tested whether or not the vertical zonation of lakes promotes idiosyncrasies in terms of community structure and diversity patterns. The mechanisms behind these patterns (environmental selection and/or allopatric speciation) were, however, difficult to be properly determined.

5.2 Methods

Because taxonomy is in a much less developed state for smaller organisms than for multicellular animals and plants, microbial ecologists have replaced "species" with other appropriate units of biodiversity based on the ribosomal RNA gene sequence, not without concerns (Rosselló-Mora & Amann, 2001) but clearly defined and consistently used in all the study systems (Ovreas, 2000; Reche et al., 2005). The slowly evolving 16S rRNA gene is the most commonly used molecular marker to survey and catalogue bacterial diversity, and it is useful for detecting ancient evolutionary events. We have recently discussed potentials and limitations of the 16S rRNA gene approach and the controversial species concept for meta-analytical and general ecology approaches on bacteria (Barberán & Casamayor, 2010). Because of the inherent biases of PCR methods used to obtain the environmental 16S rRNA gene sequences, the abundances obtained should not be strictly interpreted for absolute values but rather considered for cross-lake and cross-lake layer comparative purposes only.

5.2.1 16S rRNA gene sequence analysis

We used our own data generated from stratified lakes in Spain by genetic fingerprinting and sequencing as previously described (Casamayor et al., 2000b) and publicly available bacterial 16S rRNA gene sequences from the water col-

umn of 12 worldwide distributed stratified lakes (although 11 out of the 12 studies considered were samples from the Northern hemisphere, biasing their global perspective due to the lack of comparable studies in the Southern hemisphere, Table 5.1). The 12 lakes considered were those containing sulfide and mostly meromictic (permanent anaerobic bottom water) selected from among the many more explored by 16S rRNA gene sequencing and available in the literature. Two holomictic (hypolimnion that disappears after winter mixing) stratified lakes (i.e., Arcas and Tejo, also sulfide containing) and two deep lakes with low stratification stability (i.e., Baikal and Crater) were also included. For some of the analyses, mixolimnetic, epilimnetic, and surface layers were grouped as surface waters (SW), whereas monimolimnetic, hypolimnetic, and bottom layers of the deep lakes were analyzed as bottom waters (BW). For assessing community patterns in anaerobic hypolimnia, Lake Baikal and Crater Lake were excluded from the pool of data. Sequences shorter than 300 nucleotides were discarded, and the remaining ones were clustered at 97% identity cut-off, which is a consistent value for richness estimation (Shaw et al., 2008). The 16S rRNA gene pool was automatically aligned with the NAST aligner (DeSantis et al., 2006a), imported into ARB package (Ludwig et al., 2004), and added to a consensus phylogenetic tree with the ARB parsimony insertion tool as recently described (Auguet et al., 2010). Sequence classification into bacterial clades to assess taxonomical richness and diversity with the Shannon index followed the Ribosomal Database Project hierarchy (http://rdp.cme.msu.edu).

5.2.2 Data analysis

Phylogenetic diversity (PD) was calculated for each water layer as the sum of the branch length associated with the 16S rRNA phylogenetic tree from that layer (Faith, 1992). Distance matrices were constructed using the UniFrac metric, which is a β-diversity metric that quantifies community dissimilarity based on phylogenetic relatedness (Lozupone & Knight, 2005) and represented in a principal coordinate (PCoA) ordination plot. We performed analyses of similarity (ANOSIM) to test the hypothesis that bacterial communities from the same functional layer were more similar to each other than to communities from different layers. The ANOSIM R statistic is based on the difference of mean ranks between groups and within groups and ranges from 0 (no separation) to 1 (complete separation) (Clarke, 1993). Additionally, the distance to layer centroid was calculated as a measure of β-diversity (Anderson et al., 2006). We used standard and partial Mantel tests to determine the community concordance between UniFrac bacterial matrices from different layers and the correlation between layer-specific community matrices and ge-

Table 5.1: General characteristics of the lakes analyzed in this study. Sequences generated in this study are deposited in GenBank under accession numbers from AM749840 to AM749885.

Lake	Number of OTUs (surface)	Number of OTUs (bottom)	Method	Coordinates	Altitude (m)	Area (km^2)	Z_{max} (m)	Stratification	Trophic status
Mono (M)	20	58	Cloning	37°58′N,119°01′W	1,945	180	43	Meromictic	Eutrophic
Crater (C)	57	50	Cloning	42°56′N,122°06′W	1,882	53	589	Low	Oligotrophic
Baikal (B)	9	30	Cloning	52°00′N,107°00′E	456	31,494	1,620	Low	Oligotrophic
Kauhako (K)	12	36	Cloning	21°11′N,156°58′W	0	0.0035	248	Meromictic	Mesotrophic
Tanganyika (Ta)	20	33	Cloning	06°53′S,30°15′E	765	32,900	1,470	Meromictic	Oligotrophic
Cruz (Cr)	6	7	DGGE	39°59′N,01°52′W	1,031	0.012	24	Meromictic	Eutrophic
Tobar (To)	4	8	DGGE	40°33′N,02°03′W	1,180	0.67	19.5	Meromictic	Eutrophic
Arcas (A)	6	8	DGGE	39°59′N,02°08′W	925	0.0016	14.5	Holomictic	Eutrophic
Tejo (Te)	8	4	DGGE	39°59′N,01°52′W	1,030	0.0026	11	Holomictic	Eutrophic
Kaiike (Ki)	11	8	DGGE	30°51′N,129°52′E	0	0.15	11	Meromictic	Eutrophic
Pavin (P)	28	45	Cloning	45°55′N,02°54′E	1,197	0.44	92	Meromictic	Oligotrophic
Vilar (V)	7	5	DGGE	42°08′N,02°45′E	175	0.011	9	Meromictic	Eutrophic

Lake	Type	Ice cover	Sulfide	Anoxia	Reference
Mono (M)	Hypersaline	No	Yes	Yes	Humayoun et al. (2003)
Crater (C)	Crater	No	No	No	Urbach et al. (2001)
Baikal (B)	Rift	Yes	No	No	Glöckner et al. (2000)
Kauhako (K)	Crater	No	Yes	Yes	Donachie et al. (2004)
Tanganyika (Ta)	Rift	No	Yes	Yes	De Wever et al. (2008)
Cruz (Cr)	Karstic	No	Yes	Yes	This study
Tobar (To)	Karstic	No	Yes	Yes	This study
Arcas (A)	Karstic	No	Yes	Yes	This study
Tejo (Te)	Karstic	No	Yes	No	This study
Kaiike (Ki)	Hypersaline	No	Yes	Yes	Koizumi et al. (2004)
Pavin (P)	Crater	No	Yes	Yes	Boucher et al. (2006); Lehours et al. (2007)
Vilar (V)	Karstic	No	Yes	Yes	Casamayor et al. (2002)

ographic (S) and environmental (E) Euclidean distance matrices (Legendre & Legendre, 1998). The environmental component (E) was composed of quantitative (i.e. elevation, area, and maximum depth), semiquantitative (i.e. trophic status), and qualitative variables (i.e. stratification, geomorphology, ice cover, sulfide concentration, and anoxia; Table 5.1). Unfortunately, more detailed environmental variables were not available for the complete set of lakes studied. All statistical analyses were carried out in the R environment (http://www.r-project.org) using vegan (Oksanen et al., 2009) and picante packages (Kembel et al., 2010).

5.3 Results

We explored community patterns in terms of bacterial composition, richness, and β-diversity in the complete set of selected lakes (Table 5.1). Significant differences (p-value < 0.05, paired t-test) in bacterial community composition were found between SW and BW for major phylogenetic groups. *Actinobacteria*, *Cyanobacteria*, and *Alphaproteobacteria* were consistently more abundant in SW, while *Chlorobi*, *Gammaproteobacteria*, and *Deltaproteobacteria* dominated BW (Figure 5.1). Other groups, such as *Epsilonproteobacteria*, *Chloroflexi*, *Spirochaetes*, and *Acidobacteria*, were exclusively found in BW. Most groups abundant in SW were also found in BW, such as *Bacteroidetes*, *Verrumicrobia*, *Planctomycetes*, and *Betaproteobacteria*. The taxonomic compositions from deeper layers of Lake Baikal and Crater Lake were closer to those of the upper waters than in the case of lakes with presence of sulfide (both meromictic and holomictic, Table 5.1). Overall, the BW contained more bacterial phylogenetic groups (p-value = 0.01, paired t-test; Figure 5.2, upper panel) and were more taxonomically diverse (p-value = 0.005, paired t-test; Figure 5.2, lower panel) than the SW studied. When the diversity was analyzed taking into account the phylogenetic relationship (PD) among the different 16S rRNA gene sequences (i.e. 'tree branch length'), BW showed again greater phylogenetic diversity (PD = 33.1) than SW (PD = 22.7).

Distance matrices among environments were calculated using the UniFrac metric to explore in more detail shared phylogenetic history among the different bacterial communities. The ANOSIM analyses based on the UniFrac distance matrix showed that bacterial assemblages from the same water layer were more similar than to communities from different layer (R = 0.46, p-value < 0.001, which increased to R = 0.52, p-value = 0.002, when only meromictic lakes were considered). In turn, classification according to lake origin was not significant (R = 0.15, p-value = 0.171).

To test for β-diversity patterns, PCoA ordination plot was run using the

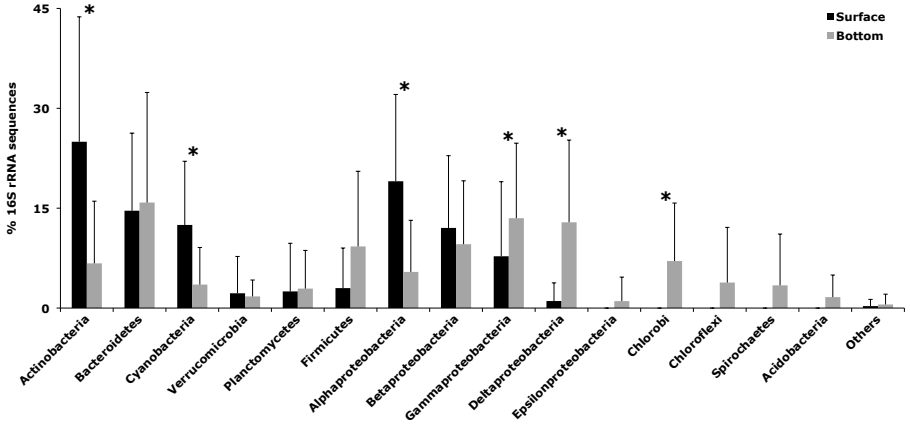


Figure 5.1: Relative abundance of bacterial taxa from the lakes listed in Table 5.1. Error bars indicate the sample standard deviation. The "Others" category consisted of members identified as *Nitrospira*, *OP11* and *OP10*. Significant differences in taxa abundance (p-value < 0.05, paired t-test) are illustrated with an asterisk.

UniFrac matrices to show ordering relationships and groups centroids for each water compartment (oxic vs. anoxic; Figure 5.3A). Anoxic hypolimnia presented greater β-diversity than oxic epilimnia measured as multivariate dispersion from the group centroid (p-value = 0.037, permutation test for homogeneity; Figure 5.3B). To assess the major factors controlling phylogenetic β-diversity patterns, the specific UniFrac matrices were compared to the same environmental (E) and spatial (S) matrices using Mantel tests. Standard Mantel tests showed significant correlations for both epilimnia (E, r_M = 0.35, p-value = 0.048; S, r_M = 0.54, p-value = 0.001) and hypolimnia (E, r_M = 0.51, p-value = 0.002; S, r_M = 0.34, p-value = 0.023). Although communities from different layers showed good concordance (r_M = 0.44, p-value = 0.005), different β-diversity patterns arose after controlling for the possible inter-matrix interactions (tested by partial Mantel tests). Surface waters showed a significant spatial component (S | E, r_M = 0.47, p-value = 0.004) and a non-significant environmental one (E | S, r_M = 0.20, p-value = 0.183). Conversely, the bottom layer was significantly correlated with environment (E | S, r_M = 0.44, p-value = 0.002) but not with geographic distance (S | E, r_M = 0.14, p-value = 0.089).

5.4 Discussion

We have used bacterioplankton populations vertically segregated in stratified lakes to determine biogeographical patterns and to suggest possible evolu-

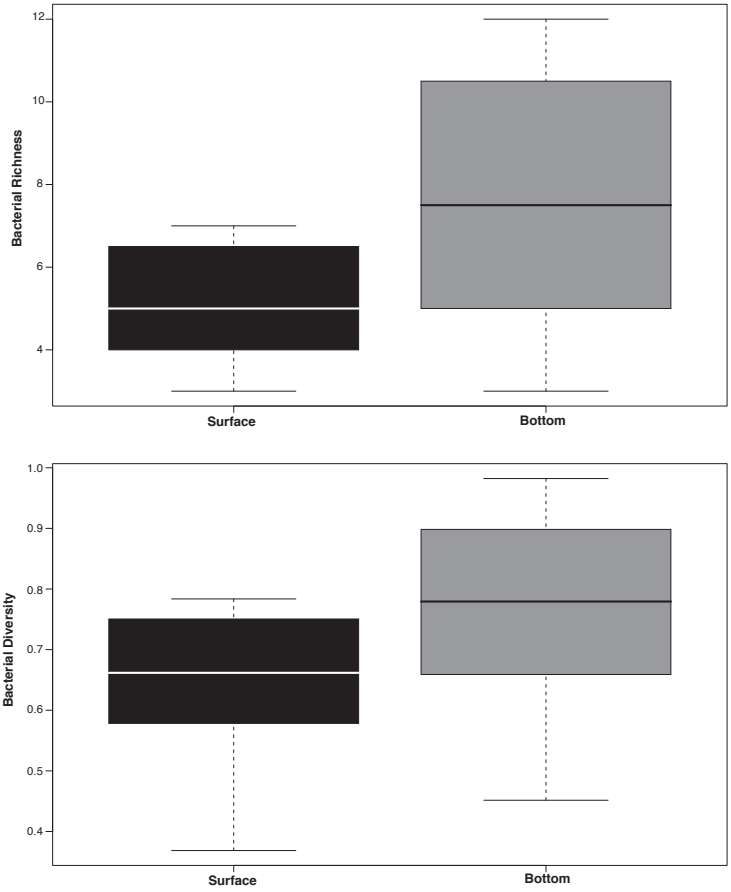


Figure 5.2: Boxplots of bacterial richness (number of taxa, *upper panel*) and bacterial diversity (Shannon diversity, *lower panel*) for surface vs. bottom waters in the different freshwater bodies studied.

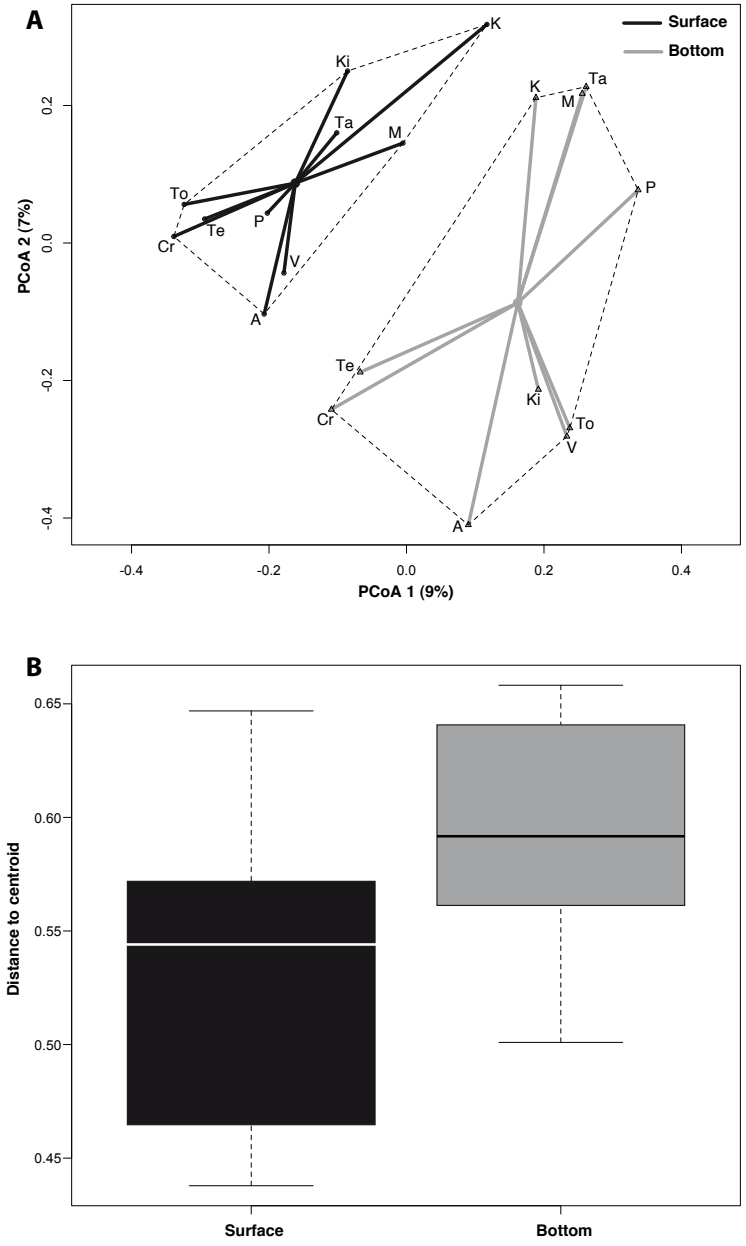


Figure 5.3: (*A*) Principal coordinates (PCoA) ordination plot with the UniFrac matrix. Lines join each sample with its corresponding layer centroid. Percentage of variation explained of the two first axes is indicated. (*B*) Boxplot of β-diversity as distance to centroid. Each point corresponds to the different freshwater bodies analyzed. Only aerobic surface waters and anaerobic bottom waters were included in this analysis.

tionary and ecological mechanisms driving microbial diversity. As earlier genetic studies indicated, these ecosystems are composed of different bacterial assemblages vertically distributed (Casamayor et al., 2002, 2000b; Konopka et al., 1999). The upper zone is dominated by taxa involved in photosynthesis (i.e. *Cyanobacteria*) or mostly in aerobic heterotrophic nutrient cycling (i.e. *Alpha-* and *Betaproteobacteria*, *Actinobacteria*). In contrast, in deeper layers, populations adapted to particular physicochemical conditions (i.e. anoxia, high sulfide concentration) such as *Gammaproteobacteria* (e.g. purple sulfur bacteria), *Deltaproteobacteria* (i.e. sulfate-respiring bacteria), or *Chlorobi* (green sulfur bacteria) were significantly more abundant. As expected, deeper layers from Lake Baikal and Crater Lake resembled surface waters due to their well-mixed nature. In addition to the particularities of each layer, most groups abundant in the epilimnion were also found in the deeper layer. This is probably due to sedimentation or sinking from upper layers that establish a downward directional gradient of passive migration (Casamayor et al., 2000b). However, different sequences were observed for the shared groups in the epi- and the hypolimnion. Thus, sinking may only partially explain the higher richness and diversity observed in deeper layers. In addition, some of the bacterial groups were exclusively found in bottom waters, and therefore, other ecological mechanisms should account for this significant difference. A plausible explanation is the biogeochemical complexity of bottom waters that creates a wide variety of functional niches to fill by those specific groups indicated above, among others. Then, bottom waters may act as a trap and also promote diversity creating rich pools of bacterial assemblages. This point is of interest to further refine the species–area relationship found for microorganisms (Reche et al., 2005, 2007), carefully considering the relationship area and niche space in these ecosystems, most of them small (< 1 km2) but with a high ratio depth/area (Table 5.1) and therefore with higher habitat heterogeneity than expected.

The comparison we carried out on different lake layers at a large scale unveiled interesting phylogenetic β-diversity patterns. In the field of community ecology, higher β-diversity values are linked to low dispersal and small connectivity between community patches (Chase, 2003). Low dispersal was expected for bottom anoxic waters for these isolated continental water bodies. Considering the temporal dynamics of a sole lake, it has been shown that the epilimnion has higher β-diversity than the hypolimnion that tends to have a more stable microbial assemblage over time (Casamayor et al., 2002; Shade et al., 2008). However, in a global spatial scale, this trend turns the opposite. Thus, lake hypolimnia promoted the formation of not only rich diverse bacterial communities but also idiosyncratic ones. The observed geographical patterns of phylogenetic β-diversity agree with the scheme of surface lay-

ers being more connected, less heterogeneous, and suffering high disturbance while bottom waters being less connected, more heterogeneous, and with low disturbance (Chase, 2003). In fact, published data indicated that anaerobic conditions prevailed for at least the last 1 million years for some of the lakes studied here (Mallorqui et al., 2005).

In stable and isolated bottom waters, endemism may be promoted by two different mechanisms. First, acting as a trap and restricting the range of the inhabitants (as described in the previous paragraph) or fostering speciation when the rate of divergence (internal novelty) exceeds dispersal rates (external novelty). Different speciation mechanisms have been proposed based on ecologically differentiated microbial populations (Koeppel et al., 2008), temporal shifts (Casamayor et al., 2002), selective sweeps (Acinas et al., 2004), resource partitioning (Hunt et al., 2008), and on variation in gene content by genome rearrangements or gene flow between populations (e.g. genomic islands; Coleman et al., 2006). Allopatric speciation results from divergent evolution of geographically isolated populations, but in microorganisms, it is hard to discern between environmental selection at the community level and allopatric speciation in geographically isolated populations (Whitaker, 2006). Isolation might occur because of great distance or a physical barrier. Data presented here suggest that, for hypolimnetic populations, both the physical barrier constituted by the permanent aerobic top water and the higher environmental heterogeneity had stronger effect than geographical distance. In turn, epilimnetic assemblages appeared more influenced by the spatial component, suggesting that high migration and distance-based stochastic effects may play a more relevant role because aerobic microorganisms would be more easily dispersed. In this line of reasoning, two different mechanisms may simultaneously be acting within and between the bacterial assemblages of the same type of ecosystems (i.e. one lake understood as one ecosystem).

Because environmental factors strongly promote stochastic events in nature, the presence/absence of particular species can constitute signals of speciation processes, and also they can reflect temporal/spatial environmental variations. Although in this study we analyzed static pictures of the freshwater bacterial assemblages, it is known that the epilimnetic communities are more dynamic and have a more marked temporal variation than their hypolimnetic counterparts (Casamayor et al., 2002). They should be, therefore, more directly submitted to environmental selection, and we should have observed higher genetic variability among epilimnia than among hypolimnia. Curiously, we detected the opposite trend, suggesting again that the combination of strong physical isolation, higher residence time, lower water inflows, and environmental heterogeneity among euxinic freshwater hypolimnia promoted diverse and rich communities with high levels of endemism.

Finally, in the present work, we studied how genetic variation (16S rRNA gene) was partitioned among geographically associated populations. Unfortunately, whether or not variations in the 16S rRNA gene sequences alone mean different bacterial "species" cannot be answered here, and deeper insights in the genome content and physiology are certainly needed (Rosselló-Mora & Amann, 2001; Whitaker et al., 2003). We have detected a general trend, but still scarce information is available on the relative importance of the interrelated ecological hierarchical factors that influence the stability and dynamics of microbial communities. This work and recent works in the literature (Galand et al., 2009, 2010; Telford et al., 2006; Whitaker et al., 2003) detected considerable regional genetic variability and, therefore, question the ubiquitous effective dispersal caused by the enormous population sizes of microbial species. Presumably, bacteria living in the anoxic hypolimnia are adapted to anaerobic metabolism and unable to survive in the oxic waters and the atmosphere for a long time. This would indicate that at least some of these bacterial species are not cosmopolitan and that endemisms do exist. We found within-lake environmental gradients that led to an efficient species sorting, whereas inter-lake dispersal capabilities were different for aerobic and anaerobic microorganisms. The scenario suggests that dispersal rates are high relative to rates of local evolution in the epilimnia, whereas the opposite trend occurs for the assemblages in the hypolimnia. Overall, endemic microbial species may appear to be more common than that previously thought in the bacterial world, and microbial ecologists should look around to find other examples of hot spots for microbial diversity. This would add to the development of a microbial diversity conservation perspective still lacking in the global and local biodiversity preservation policies.

Acknowledgements

We are thankful to authors who provided valuable data for the analysis and to two anonymous reviewers for their constructive comments. This research was supported by grant AERBAC 079/2007 from the Spanish Ministerio de Medio Ambiente (MARM) and grants CONSOLIDER-INGENIO 2010 GRACCIE CSD2007-00067 and PIRENA CGL2009-13318-CO2-01/BOS from the Spanish Ministerio de Ciencia e Innovación (MICINN) to EOC. AB is supported by the Spanish FPU predoctoral scholarship program.

6

Global ecological patterns in uncultured *Archaea*

Resumen

Al aplicar una aproximación filogenética global a organismos *Archaea* no cultivados se han revelado patrones de comunidades definidos a lo largo de amplios gradientes y tipologías ambientales. El análisis se ha basado en unas 2000 secuencias del gen 16S ribosomal de arqueas provenientes de 67 localizaciones. Dichas secuencias fueron agregadas al 97 % de identidad, clasificadas en siete tipos de hábitats, y analizadas tanto con UniFrac (para explorar la historia filogenética compartida) como mediante árboles de regresión multivariantes (que consideran la abundancia relativa de los diferentes linajes). Ambas perspectivas apuntaron a la salinidad como el factor regulador principal a escala global. Las chimeneas hidrotermales y los hábitats planctónicos continentales se postularon como los mayores reservorios de diversidad de arqueas y, por lo tanto, como ambientes prometedores para el descubrimiento de nuevos linajes. Por el contrario, los suelos exhibieron una mayor agregación filogenética, resultado de la presencia de filotipos relacionados estrechamente. Se detectaron diferentes linajes indicadores para los distintos hábitats, algunos de los cuales desconocidos a nivel ecológico. Según los análisis de distribución de comunidades, las chimeneas hidrotermales parecen ser uno de los primeros hábitats colonizados por arqueas. En resumen, este estudio proveyó de soporte ecológico a la nomenclatura arbitraria de *Archaea*, a la par que desveló aspectos filogeográficos relevantes en su biología.

Abstract [1]

We have applied a global analytical approach to uncultured *Archaea* that for the first time reveals well-defined community patterns along broad environmental gradients and habitat types. Phylogenetic patterns and the environmental factors governing the creation and maintenance of these patterns were analyzed for ca. 2,000 archaeal 16S rRNA gene sequences from 67 globally distributed studies. The sequences were dereplicated at 97% identity, grouped into seven habitat types, and analyzed with both UniFrac (to explore shared phylogenetic history) and multivariate regression tree (that considers the relative abundance of the lineages or taxa) approaches. Both phylogenetic and taxon-based approaches showed salinity and not temperature as one of the principal driving forces at the global scale. Hydrothermal vents and planktonic freshwater habitats emerged as the largest reservoirs of archaeal diversity and consequently are promising environments for the discovery of new archaeal lineages. Conversely, soils were more phylogenetically clustered and archaeal diversity was the result of a high number of closely related phylotypes rather than different lineages. Applying the ecological concept of "indicator species", we detected up to 13 indicator archaeal lineages for the seven habitats prospected. Some of these lineages (that is, hypersaline MSBL1, marine sediment FCG1 and freshwater plSA1), for which ecological importance has remained unseen to date, deserve further attention as they represent potential key archaeal groups in terms of distribution and ecological processes. Hydrothermal vents held the highest number of indicator lineages, suggesting it would be the earliest habitat colonized by *Archaea*. Overall, our approach provided ecological support for the often arbitrary nomenclature within uncultured *Archaea*, as well as phylogeographical clues on key ecological and evolutionary aspects of archaeal biology.

6.1 Introduction

The study of the biology and ecology of *Archaea* is currently among the most exciting and dynamic research topics in microbial ecology. In less than two decades the status of these enigmatic microorganisms has changed completely. The popularization of environmental ribosomal gene analysis has revolutionized the biased perception on their biology and ecology. The new tools have expanded archaeal ecological distribution and metabolic diversity far beyond expected, unveiling a widespread distribution and an unexpected

[1] See original publication in Auguet et al. (2010).

diversity (Schleper et al., 2005; Chaban et al., 2006; Auguet & Casamayor, 2008; Llirós et al., 2008; Casamayor & Borrego, 2009).

The earliest archaeal phylogenetic tree derived from laboratory cultures (hyperthermophiles, halophiles and methanogens) was composed of the two main phyla, *Crenarchaeota* and *Euryarchaeota*, and contained a few branches. However, environmental PCR-based 16S rRNA gene surveys quickly expanded the archaeal tree with the discovery of new uncultured lineages. One of the most noticeable advances during the nineties was the discovery of mesophilic *Crenarchaeota* inhabitants of marine plankton and soils that formed a deeply divergent clade distantly related to hyperthermophiles. Two main crenarchaeal lineages were observed within this new clade: the 1.1a (DeLong, 1992; Fuhrman et al., 1992) and the 1.1b (Bintrim et al., 1997; Ochsenreiter et al., 2003). In the last years, the 16S rRNA gene sequences from uncultured *Archaea* in databases have increased several orders of magnitude above those available from the cultured counterparts. A precise taxonomic placement of the new sequences will remain, however, uncertain until microbiologists succeed bringing into culture more archaeal representatives from a larger range of phyla. In addition, almost half of the 16S rRNA gene sequences archived in GenBank database lack clear taxonomic information (DeSantis et al., 2006b). As a consequence, different authors use different names for uncultured clusters that lead to conflicting nomenclatures, and ecological or physiological information becomes often veiled behind confusing clusters naming.

At present, public databases hold a large number of archaeal 16S rRNA environmental sequences (ca. 40,000) from a large set of environments. This data set contains information to extract general macroecological patterns and to bring some light on how archaeal communities are structured along global environmental gradients. The aims of this study are to use the information present in databases to (i) describe the global distribution of archaeal communities and understand the forcing environmental factors that shape archaeal diversity and (ii) detect the main taxa that can be considered as "indicator species" for a given habitat. We also provided a framework to identify environments that contain the highest archaeal diversity and represent promising habitats for the discovery of new archaeal lineages.

6.2 Methods

6.2.1 Construction of the archaeal 16S rRNA gene database

We surveyed published literature and GenBank database for archaeal 16S rRNA clone libraries (that is, a collection of identified PCR products obtained from the same source) that matched each one of the following criteria: (i) communities obtained from natural environments (artificial and semi-artificial environments with human-induced dynamics, such as rice soils and chemical reactors, were excluded for detailed analyses. In fact when sorted into an ordination plot according to phylogenetic community similarity, rice soils significantly separated from typical natural soil environments and were closer to freshwater sediments [data not shown]); (ii) high-quality data (no nucleotide ambiguities present and sequences > 300 bp); and (iii) use of universal primers covering the same 16S rRNA gene region. We homogenized different methodologies and sampling efforts by clustering sequences at 97% identity threshold (Shaw et al., 2008). We ended with an archaeal database of ca. 2,000 archaeal 16S rRNA sequences from 67 clone libraries globally distributed [2]. The sequences were treated by two methods (see below), that is, by using an explicitly phylogenetic approach, and by a taxon-based approach (where taxa were picked at a defined level and then treated as equally divergent).

The different clone libraries were grouped into seven distinct habitats (understood as a group of environments sharing a close geochemistry) as follows: freshwater plankton (Fwc), freshwater sediment (Fsed), soil (S), marine plankton (Mwc), marine sediment (Msed), hypersaline planktonic environments (Hsal) and hydrothermal vents (Hdv). Next, we constructed a semiquantitative environmental matrix according to the range of environmental gradients present in these habitats: temperature (hydrothermal vents to polar waters), salinity (hyperhaline brines to freshwater), life environment (plankton, soil and sediment), trophic state (eutrophic to ultraoligotrophic) and oxygen concentrations (anoxic to full oxic).

The 16S rRNA gene sequences were automatically aligned with the NAST aligner (DeSantis et al., 2006a) and imported into the Greengenes database (DeSantis et al., 2006b) based on the ARB package (Ludwig et al., 2004). A base frequency filter was applied to exclude highly variable positions before sequences were added using the ARB parsimony insertion tool to the original Greengenes tree calculated by maximum parsimony method and provided by default.

[2] See Supplementary Information in Auguet et al. (2010).

6.2.2 Phylogenetic approach

Distance matrices were constructed using the UniFrac metric. UniFrac is a beta diversity metric that quantifies community similarity based on the phylogenetic relatedness (Lozupone & Knight, 2005). To assess the sources of variation in the UniFrac matrix, we used Permutational Multivariate ANOVA (PERMANOVA) based on 1,000 permutations (McArdle & Anderson, 2001) with function adonis in vegan R package (Oksanen et al., 2009).

Phylogenetic diversity (PD) for each of the seven habitats was calculated as the sum of the branch length associated with the 16S rRNA gene sequences within this habitat (Faith, 1992). To correct for unequal number of sequences, we calculated the mean PD of 1,000 randomized subsamples of each habitat (Barberán & Casamayor, 2010).

The phylogenetic structure for each habitat was calculated with the phylogenetic species variability (PSV) index (Helmus et al., 2007). PSV quantifies how phylogenetic relatedness decreases the variance of a hypothetical neutral trait. The value is 1 when all species are phylogenetically unrelated (that is, a star phylogeny) and approaches 0 as species become more related. To statistically test whether habitats were composed of species that are more or less related to each other than expected, we compared the mean observed PSV with distributions of mean null values (1,000 iterations) using two different randomization procedures. Null model 1 maintains species occurrence, whereas null model 2 maintains habitat species richness (Helmus et al., 2007). Analyses were run with the R package picante (Kembel et al., 2010).

A genetic distance matrix of the sequences from each habitat was constructed with a subset of studies that amplified the same 16S rDNA region. This matrix was imported to DOTUR (Schloss & Handelsman, 2005) and used to determine OTUs and to calculate rarefaction curves.

6.2.3 Taxon-based approach

Archaeal lineages were named following the clusters or divisions naming immediately subordinate to the *Crenarchaeota* or *Euryarchaeota* phyla and provided by default in the Greengenes tree. However, several sequences seemed not related to any labeled cluster and would have remained unaffiliated at the lineage level. Accordingly, we named four new crenarchaeotal lineages de novo as follows: 1.1d, 1.1e, 1.1f and 1.1 g, and one euryarchaeotal lineage as HV-Fresh (see Figure 6.4). The HV-Fresh lineage not only contained the already described DHVE3 (Deep Hydrothermal Vent group 3) and HV1 (Hydrothermal Vent group 1), but also a large number of single freshwater sequences. Grouping at a lower phylogenetic level was ruled out because of the

high number of archaeal sequences not properly affiliated yet and the poor taxonomic agreement due to the lack of cultured representatives.

Microbes have a great capacity of dispersion and one sequence of any lineage can be retrieved in any habitat by chance. Furthermore, cross-contamination is very possible when sampling at the interface of two habitats (for example, sediment-water column). Hence, to identify archaeal lineages as analogous to the concept of "indicator species" for each habitat with enough statistical support, we constructed a table of abundances and used the indicator value (IndVal) index, which combines relative abundance and relative frequency of occurrence (Dufrene & Legendre, 1997). A multivariate regression tree was computed with the R package mvpart (De'Ath, 2002) in order to represent the relationship between the table of lineage abundances and the environmental matrix.

6.3 Results

6.3.1 Environmental forces shaping the phylogenetic structure of archaeal community

Natural samples from 67 globally distributed archaeal clone libraries were sorted into an ordination plot according to phylogenetic community similarity (Figure 6.1a). Habitat classification was a strong structuring factor of the archaeal assemblages ($R^2 = 0.20$, p-value < 0.001) and communities grouped according to their habitat of origin (Figure 6.1a). Nonsaline environments clearly separated from saline environments (Figure 6.1), and salinity was the strongest and the only significant environmental factor ($R^2 = 0.03$, p-value = 0.024). The remaining environmental factors explored (that is, temperature, life environment, oxygen concentration and trophic status) were not significant and explained only 8.4% of the total variance from the UniFrac matrix.

Rarefaction curves (Figure 6.2) and diversity indices (Figure 6.1b) were determined for the seven types of habitats. The linear rarefaction curves provided evidences that the archaeal diversity is far from exhaustively sampled, particularly in freshwater, hydrothermal vent and hypersaline habitats. PD was higher in freshwater plankton (Fwc) and hydrothermal vents (Hdv), whereas soil (S) hold the lowest PD value (Figure 6.1b). Habitats showed a nonrandom sampling of phylotypes from the phylogeny pool, thereby indicating a significant phylogenetic structure. The mean observed PSV value (0.53) was significantly lower than the null distribution for model 1 (0.75, p-value < 0.05) and for model 2 (0.60, p-value < 0.05). Null model 1 test suggested nonrandom associations between phylotypes among communities,

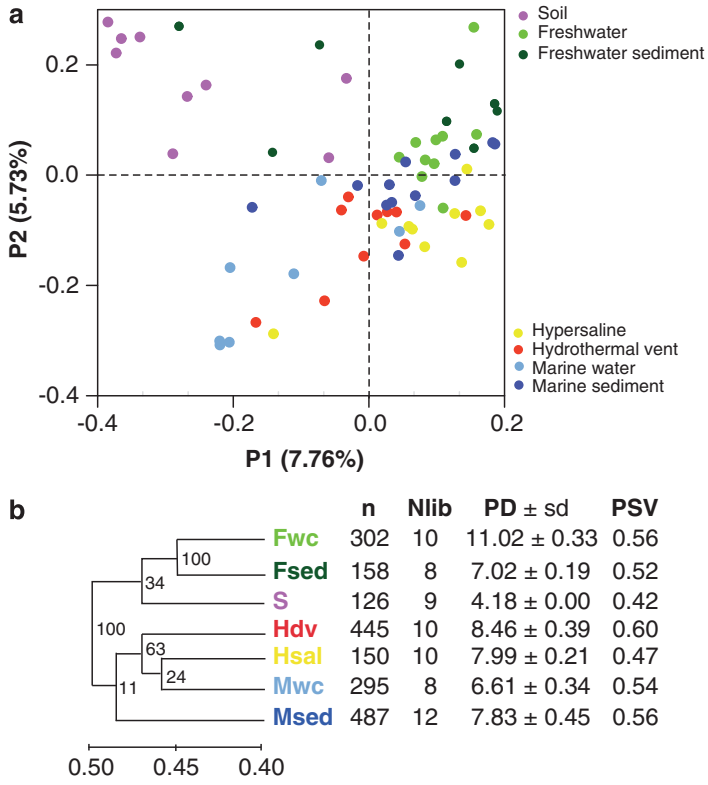


Figure 6.1: (*a*) Principal coordinates analysis (PCoA) obtained with the UniFrac distance matrix comparing the 67 libraries. (*b*) Hierarchical clustering analysis (UPGMA algorithm with Jackknife supporting values, 126 subsampled sequences, 100 replicates) carried out on the libraries belonging to the seven habitats type previously defined by the PCoA analysis. Distances between clusters are expressed in UniFrac units: a distance of 0 means that two environments contain mutually exclusive lineages. The number of sequences (n), number of libraries (Nlib), phylogenetic diversity with s.d. (PD±s.d.) and phylogenetic species variability (PSV) in each habitat is given. S.d. for PSV index was < 0.001 for all habitats.

with habitats containing more closely related phylotypes than expected by chance (that is, phylogenetic clustering). The null model 2 suggested that phylotype composition represented nonrandom samples from the phylotypes pool (that is, significant pattern in phylotypes prevalence). Particularly, Hdv, Fwc and Msed habitats showed the highest PSV values (that is, more overdispersed), whereas S and Hsal the lowest (that is, more phylogenetically clustered).

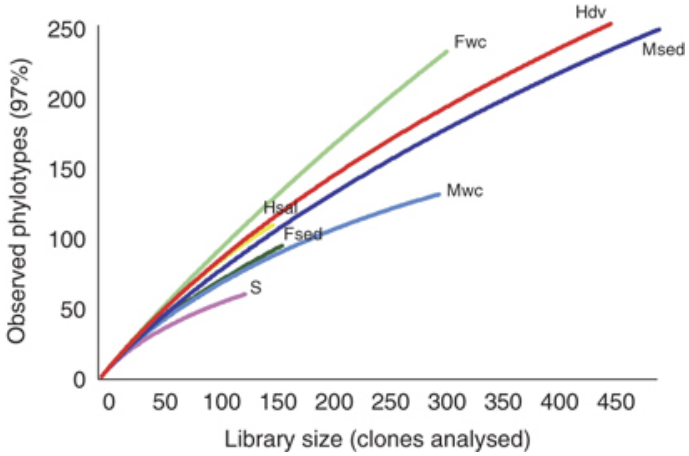


Figure 6.2: Rarefaction curves for archeal diversity in the seven habitats prospected. OTUs were calculated at a 97% cut-off.

6.3.2 Identifying indicator lineages and their distribution along gradients

Both the Shannon index and the richness values (Figure 6.3 upper part) showed again that hydrothermal vents (Hdv) and freshwater plankton (Fwc) were the most diverse habitats, whereas soil (S) was the lowest. Overall, 13 out of 25 archaeal lineages showed a significant IndVal (p-value < 0.01) for one single habitat (labeled with asterisk in Figure 6.3). The 1.1b, FCG1, *Thermococcales*, and *Thermoproteales* had high IndVal values (range: 63–90), whereas the remaining lineages showed moderate values (range: 30–49). *Methanomicrobiales* and *Thermoplasmatales* were predominant in freshwater and marine habitats, respectively, but the analysis was not significant for any of them (p-value > 0.01). Freshwater archaeal communities were dominated by the indicator archaeal group plSA1. Soil samples were dominated by the crenarchaeal lineages 1.1b and 1.1c (abundance 76±33 %). Conversely, *Crenarchaeota* were essentially absent from hypersaline samples, where *Euryarchaeota* from the *Halobacteriales*+SA1 lineages dominated. Remarkably, almost all the phylogenetic groups were present in hydrothermal vents with some specific groups exclusively found there. Hydrothermal vents also showed by far the largest number of indicator lineages (five lineages), most of them located close to the root of the tree (Figure 6.4). Curiously, for freshwater sediments none of the lineages were detected as indicator at 0.01 significance level, though *Methanomicrobiales* became significant at p-value < 0.05.

A multivariate regression tree analysis was carried out in order to link

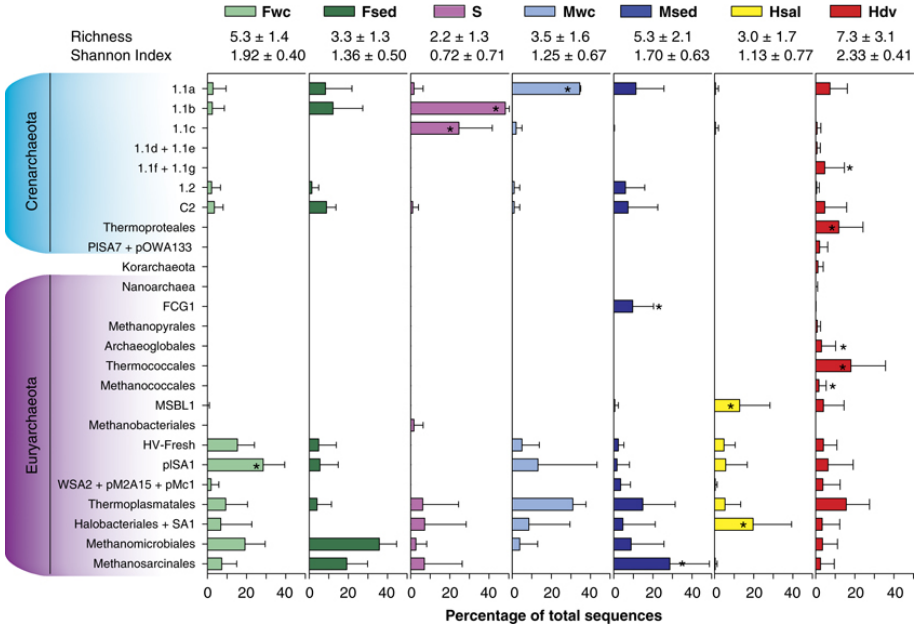


Figure 6.3: Relative proportion of archaeal lineages (based on sequence abundance) within each of the seven habitats identified. The number of libraries and sequences used for each habitat is given in Figure 6.1b. Error bars represent s.d. Asterisks show indicator archaeal lineages at a significance threshold of p-value = 0.01. The richness and Shannon diversity index are given in the upper part of the figure for each habitat.

the abundance of the lineages and environmental data. The analysis showed an eight-leaf tree ordination (Figure 6.5) primarily based on life environment (soils vs. sediment and plankton), and followed by salinity (hypersaline vs. marine and freshwater), oxygen level and temperature. The ordination explained 38.5% of the phylogenetic lineage variance. As previously observed for UniFrac analyses, samples clustered in the leaves of the tree merely in function of their habitat of origin. Nonetheless, some samples from related habitats grouped together forced by other environmental parameters. Thus, anoxia tended to pool together Hdv and Msed (hot- and cold-temperate anoxic marine sites), as well as Fsed and Fwc (sediments and water column from anoxic freshwaters), whereas Hsal environments were separated between oxic and anoxic (Figure 6.5).

Pie charts in Figure 6.5 show in detail how the relative abundance of each phylogenetic group contributed to the separation and composition of the leaves. Indicator lineages previously identified by the IndVal index were mainly responsible for the regression tree topology observed. For example,

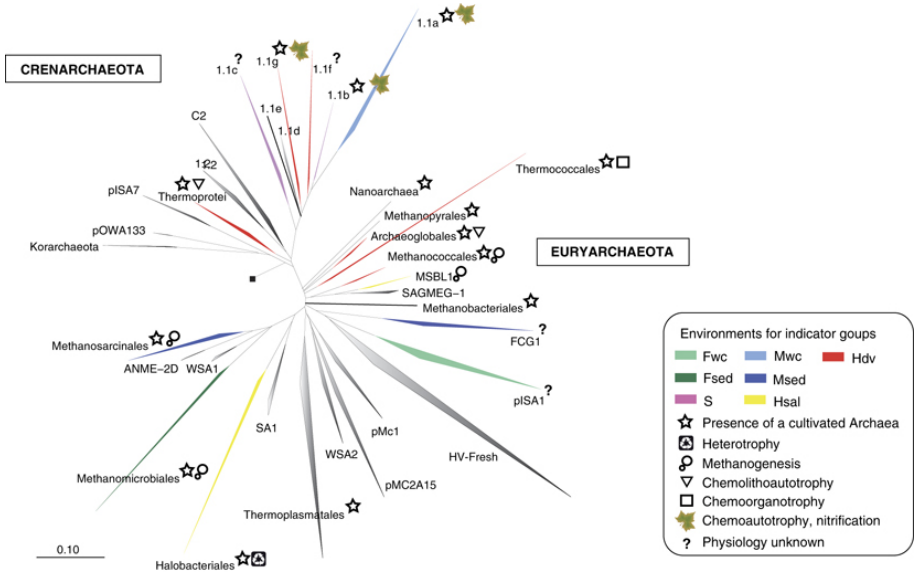


Figure 6.4: Phylogenetic archaeal tree based on 16S rRNA gene sequences present in the Greengenes database for the ARB software in January 2008. Sequences were inserted into the original Greengenes tree calculated by maximum parsimony method and provided by default by using parsimony criteria with the *Archaea* filter excluding highly variable positions. The nomenclature follows the labeled clusters or divisions provided by default in the Greengenes general tree and immediately subordinate to the *Crenarchaeota* or *Euryarchaeota* phyla. The black square in the centre indicates rooting to species in the domain *Bacteria*. Only the physiologies of indicator lineages were represented.

the crenarchaeotal lineages 1.1b and 1.1c largely determined the initial separation soil vs aquatic environments in agreement with the previous analyses that classified these lineages as typical soil inhabitants.

6.4 Discussion

Archaeal ecology derived from cultured representatives (54 cultured species reported so far, and spread in 18 lineages, Schleper et al., 2005) provided for many years a strongly biased view on the diversity and distribution of the third Domain of life in the biosphere. Basically, they were considered extremophiles thriving under severe extremes of environmental gradients. The emergence of culture independent molecular techniques unveiled the ubiquitous distribution of *Archaea* (Casamayor & Borrego, 2009; Chaban et al., 2006; Schleper, 2007) and also revealed a hidden PD in the Domain with, to date,

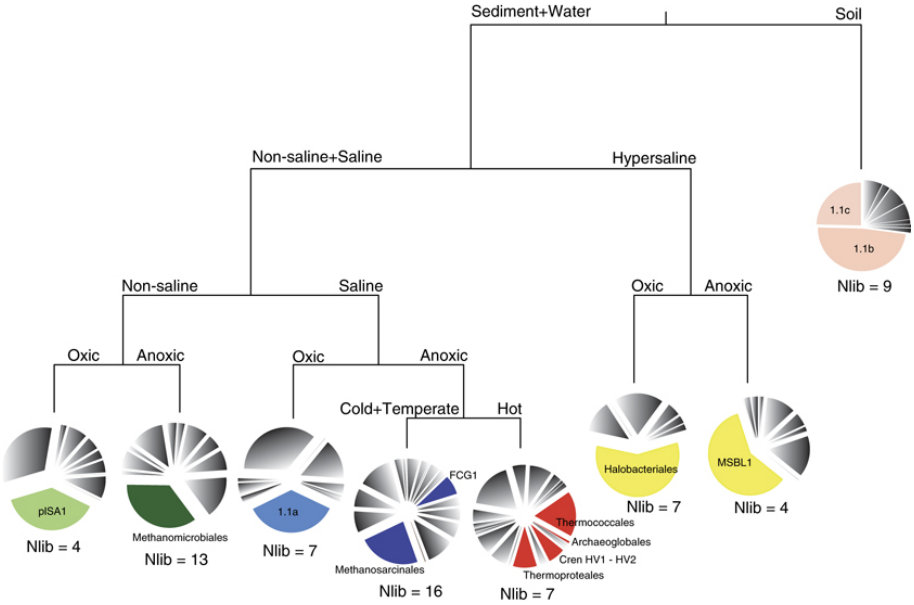


Figure 6.5: Multivariate regression tree (MRT) analysis of the interaction between archaeal lineage abundance (in terms of sequence number) and environmental parameters. The model explained 38.5% of the variance in the whole data set. Pies under each leaf represent the mean of normalized archaeal lineage abundance for each lineage significantly correlated with environmental parameters.

up to 49 mostly uncultured lineages (Schleper et al., 2005; Schleper, 2007). Obviously, to gain knowledge on the true ecology of the Domain, all the components should be analyzed as a whole. However, detailed comparative ecological studies to fully appreciate the distribution, community patterns and environmental drivers of uncultured *Archaea* are missing. To fill this gap, our analytical approach revealed for the first time well-defined community patterns along global environmental gradients and habitat types for uncultured *Archaea*.

Archaeal communities were more similar within habitats than among habitats. This clustered phylogenetic structure (that is, more closely related phylotypes than expected from a random distribution within habitats) is consistent with the concept of habitat filtering (Helmus et al., 2007). Curiously, salinity rather than temperature explained a significant part of these distribution patterns. Salinity was also recently recognized as a key environmental factor globally structuring bacterial communities (Lozupone & Knight, 2007). Other environmental factors such as oxic–anoxic conditions probably also

had a significant role structuring the observed patterns. Overall, the two phylogenetically independent domains of life (that is, *Archaea* and *Bacteria*) shared similar broad trends, suggesting a commonality in the types of factors that are important for prokaryotes distribution.

The lack of environmental information associated with database sequences was, however, a strong limitation of meta-analysis and may have hampered a better explanation for the global archaeal patterns observed here. As stated by other authors, it becomes crucial to gain a consensus rationale for measuring and reporting some basic environmental variables in microbial surveys (Robertson et al., 2005; Field et al., 2008). In addition, our approach contained a stochastic component inherent to any environmental study (Sloan et al., 2006). And the 16S rRNA molecule may not be the most suitable marker to target fine biogeographical patterns because of its highly conserved nature. However, several studies agree that this approach remains still valuable as the first step for exploring general ecological patterns in uncultured microorganisms (Reche et al., 2005, 2007; Ramette & Tiedje, 2007, and references therein).

Nonetheless, these limitations were not strong enough to blur the powerful effect of local environmental selection on archaeal diversity. Interestingly, both phylogenetic and taxon-based approaches revealed similar diversity patterns, suggesting that all the phylotypes of a lineage roughly shared the same distribution and probably the same physiology as in the case of methanogens, halophiles, thermophiles and ammonia oxidizers, all of them clustered in specific functional groups. These observations agree with a recent study (Von Mering et al., 2007) that showed a significant correlation between habitat and evolutionary relatedness for microorganisms, even for taxa related at the order level, suggesting that truly adapted specialists acquired their abilities long time ago. To ascertain who were the true archaeal specialists, we use the concept of indicator species borrowed from plant and animal ecology.

We defined as specialists those archaeal lineages that were more frequently represented in most of the sites within a specific habitat, a definition closely related to the concept of indicator species used in ecology (Dufrene & Legendre, 1997). We applied this concept to the whole set of archaeal lineages, unveiling up to 13 lineages with significant IndVal values. We found at least one indicator lineage for each habitat at a significance threshold of p-value = 0.01, except for freshwater sediments. Thus, this original approach provided a novel ecological support for the sometimes arbitrary nomenclature found in uncultured archaeal clusters, and heavily supported the attributes previously given for crenarchaeotal 1.1a as the marine planktonic group ("marine plankton group 1", DeLong 1998), or to the 1.1b as the soil crenarchaeotal group.

In addition, this approach provided new phylogeographical clues on ecological and evolutionary aspects of the archaeal biology. As acquisition of

the essential functions to be permanently adapted to a habitat requires an extended period of time (Von Mering et al., 2007), the lack of indicator archaeal lineages in freshwater sediments may indicate a late archaeal colonization. Conversely, hydrothermal vents exhibited by far the highest number of indicator lineages and this may indicate that we were dealing with the earliest habitat colonized by *Archaea*. Although the thermophilic origin of planktonic *Archaea* is still a matter of debate (DeLong, 1998; Brochier-Armanet et al., 2008), the presence in hydrothermal vents of representatives from almost all archaeal lineages (especially those at the root of the tree in Figure 6.4) offers another piece of the puzzle supporting Hdv as the cradle of planktonic *Archaea* and probably of the origin for the common archaeal ancestor.

The indicator archaeal lineages identified by the IndVal index produced the clustering topology observed in the multivariate regression tree analysis, confirming that these groups were the best-adapted assemblages to the prevailing environmental conditions for each habitat. If environmental forcing selects microorganisms on the basis of their functional capacities, indicator lineages should be, consequently, among the main players in the pivotal ecological functions within the habitat. Good examples are the *Methanomicrobiales* and *Methanosarcinales* (indicator groups for freshwater and marine sediments, respectively), well known as central components for anaerobic organic matter degradation coupled to methanogenesis in aquatic environments. *Haloarchaeales* also constitute the most active population for organic matter degradation in hyperhaline environments (e.g. Gasol et al., 2004), whereas in hydrothermal vents chemolithoautotrophs members of the *Archaeoglobales* and the *Thermoprotei* are recognized as key primary producers under anaerobic conditions by coupling oxidation of hydrogen gas with sulfate reduction (Segerer et al., 1993, and references therein). Furthermore, from the recent cultivation of the autotrophic ammonia oxidizer *Crenarchaeota Nitrosocaldus yellowstonii* (De la Torre et al., 2008), we can hypothesize a significant role in the nitrogen cycle of Hdv by members of the 1.1g group. Similarly, *Crenarchaeota* from the 1.1a and 1.1b groups are thought to be important nitrifiers in planktonic marine systems and soils (Francis et al., 2007). Finally, 3 out of 13 indicator archaeal lineages (that is, MSBL1, FCG1 and plSA1) did not contain cultivated counterparts or genomic fragments good enough to extract functional information (Figure 6.4). For the MSBL1 lineage, however, a putative methanogenic metabolism was inferred according to its phylogenetic allocation (van der Wielen et al., 2005). No physiological information is available so far for the FCG1 lineage (marine sediments) and, particularly, for the plSA1 lineage (characteristic of freshwater habitats) where peculiar fast evolving 16S rRNA gene sequences (long branches) were observed (Figure 6.4). These three lineages, for which ecological importance has remained unseen

to date, deserve further and detailed attention as they represent potential key archaeal groups in term of distribution and ecological processes in their respective habitats.

In the coming future, new genomic tools will offer a wider picture of the archaeal diversity that will probably lead to substantial changes in current archaeal phylogeny (Brochier-Armanet et al., 2008; Robertson et al., 2005; Schleper et al., 2005). A correct positioning of the lineages within the phylogenetic tree topology is a fundamental issue to extract insights into the evolution and metabolic capacities of uncultured *Archaea*. In essence, both success in bringing into culture more archaeal representatives from a larger range of phyla and a higher sequencing effort are still needed to get a more realistic picture of archaeal diversity and phylogeny. In this context, the diversity analyses reported here offered unexpected views unveiling hydrothermal vents and planktonic freshwater ecosystems as the largest reservoirs of archaeal diversity and, therefore, promising environments for the discovery of new archaeal lineages. This would encourage a new focus on sequencing efforts, as these two habitats are by far less thoroughly sampled than soil or marine habitats (for example, planktonic freshwater sequences only represent 2.5% of total archaeal sequences in GenBank). Conversely, archaeal diversity in soils was unexpectedly low even though soil microbial diversity is assumed to be one of the highest on Earth (Torsvik & Ovreas, 2002). As previously observed for soil bacteria (Lozupone & Knight, 2007), the archaeal soil diversity was the result of a high number of closely related phylotypes rather than different lineages. This is confirmed here by the lowest PSV value indicating a high degree of phylogenetic clustering in soil archaeal assemblages. This peculiar characteristic of soils as compared with aquatic habitats certainly deserves further attention. A first element to consider would be the lowest evolutionary rates observed in soils that could be related to the faculty of microorganisms to enter in dormancy during long stressing periods (for example, winter, desiccation; Von Mering et al., 2007).

Overall, our approach revealed for the first time well-defined global patterns in the distribution of uncultured *Archaea* with a strong environmental filtering component. Archaeal indicator lineages were identified for specific habitats leading the classification of uncultured *Archaea* into a more comprehensive and ecological framework. Such lineages appear as good targets in future research for finely depicting the links between ecological drivers and archaeal biology. Emerging patterns will help to guide future research on archaeal biology and ecology.

Acknowledgements

We are thankful to all the authors who provided valuable data for this work. We also acknowledge anonymous reviewers for valuable feedbacks and constructive comments. This research was supported by grant CRENYC CGL2006-12058 to EOC from the Spanish Ministerio de Educación y Ciencia (MEC), and CONSOLIDER-INGENIO 2010 project GRACCIE CSD2007-00004. JCA benefits from a Juan de la Cierva-MEC postdoctoral fellow, and AB is supported by an FPU-MEC predoctoral scholarship.

7

Phylogenetic ecology of widespread uncultured clades of the Kingdom *Euryarchaeota*

Resumen

A pesar de su extensa distribución y su gran diversidad filogenética, los microorganismos continúan siendo desconocidos a nivel ecológico. Con el fin de ahondar en la distribución ambiental y la historia evolutiva se aplicó una aproximación de ecología filogenética basada en secuencias del gen 16S rRNA a dos grupos del reino *Euryarchaeota*, Lake Dagow Sediment (LDS) y Rice Cluster-V (RC-V). La historia evolutiva inferida indicó que ambos grupos han sufrido evolución específica en cada ambiente, con notables eventos de transición entre hábitats. Comparado con otros grupos microbianos de arqueas, ambos grupos presentaron remarcables niveles de diversidad genética posiblemente fomentada por su adaptabilidad ambiental y la heterogeneidad de las masas de agua continentales donde medran. Los diversificación a lo largo de la historia filogenética se concentró tanto en los instantes iniciales como en los más recientes. Para la mayoría de microorganismos, la diferenciación genética y fisiológica que habilita la explotación de los recursos ambientales sigue desconocida. La inferencia de la historia evolutiva a partir de filogenias moleculares basadas en el gen ribosomal del 16S permite una perspectiva ecológica con la que escudriñar las estrechamente entrelazadas relaciones entre los linajes, el ambiente y la contingencia histórica en el mundo microbiano.

Abstract [1]

Despite its widespread distribution and high levels of phylogenetic diversity, microbes are poorly understood creatures. We applied a phylogenetic ecology approach in the Kingdom *Euryarchaeota* to gain insight into the environmental distribution and evolutionary history of one of the most ubiquitous and largely unknown microbial groups. We compiled 16S rRNA gene sequences from our own sequence libraries and public genetic databases for two of the most widespread mesophilic *Euryarchaeota* clades, Lake Dagow Sediment (LDS) and Rice Cluster-V (RC-V). The inferred population history indicated that both groups have undergone specific non-random evolution within environments, with several noteworthy habitat transition events. Remarkably, the LDS and RC-V groups had enormous levels of genetic diversity when compared with other microbial groups, and proliferation of sequences within each single clade was accompanied by significant ecological differentiation. Additionally, the freshwater *Euryarchaeota* counterparts unexpectedly showed high phylogenetic diversity, possibly promoted by their environmental adaptability and the heterogeneous nature of freshwater ecosystems. The temporal phylogenetic diversification pattern of these freshwater *Euryarchaeota* was concentrated both in early times and recently, similarly to other much less diverse but deeply sampled archaeal groups, further stressing that their genetic diversity is a function of environment plasticity. For the vast majority of living beings on Earth (i.e. the uncultured microorganisms), how they differ in the genetic or physiological traits used to exploit the environmental resources is largely unknown. Inferring population history from 16S rRNA gene-based molecular phylogenies under an ecological perspective may shed light on the intriguing relationships between lineage, environment, evolution and diversity in the microbial world.

7.1 Introduction

The largest part of microorganisms is still uncultured, and therefore, their ecology, physiology and metabolic potentials remain widely unknown. The relationships between lineage, environment, evolution and diversity are also largely unexplored in the microbial world. Over the last 15 years, very few microorganisms have attracted so much interest as the *Archaea* (mainly the Kingdoms *Crenarchaeota* and *Euryarchaeota*), the third domain of life. These widespread enigmatic microbes have changed the microbial ecologists' point of view on archaeal ecology with the discovery of mesophilic archaea (mostly

[1] See original publication in Barberán et al. (2011).

Crenarchaeota) as abundant members of aquatic ecosystems (e.g. Karner et al., 2001; Auguet & Casamayor, 2008) and key components linking the global carbon (i.e. carbon dioxide fixing, Auguet et al., 2008; De la Torre et al., 2008; Herndl et al., 2005) and nitrogen (i.e. ammonia oxidation, Francis et al., 2005) cycling. Most of the knowledge currently available for mesophilic *Crenarchaeota* relies both on molecular in situ studies (surveys on the 16S rRNA gene and a few functional markers such as the crenarchaeotal amoA gene that codes for an ammonia monooxygenase) and on the successful culturing of the marine mesophilic *Crenarchaeota* Candidatus *Nitrosopumilus maritimus* (Könneke et al., 2005). What is currently known on the ecophysiological relevance of *Euryarchaeota* mainly arises from the strains grown in pure culture and on the ca. 30 genomes available in databases. Most of them belong, however, to extremophiles (i.e. thermophiles and halophiles) and methanogens, distantly related to their mesophilic counterparts. There is therefore a big gap in the current knowledge of the ecology and population history of mesophilic *Euryarchaeota* beyond methanogens and halophiles.

Despite the lack of information on the whole ecological potential of mesophilic *Euryarchaeota*, the increasing number of environmental ribosomal gene surveys has expanded their phylogenetic scope with a large number of uncultured lineages. Mesophilic *Euryarchaeota* have been initially described in marine environments where they mainly belong to Group II, III and IV within the *Thermoplasmatales* lineage, Group II being the most abundant (DeLong, 1998; Auguet et al., 2010; Galand et al., 2010). Although less documented, uncultured *Euryarchaeota* have also been frequently recovered from inland waters and essentially belong to two highly diverse lineages originally called LDS (Lake Dagow sediment, Glissman et al., 2004) and RC-V (Rice Cluster-V, Grosskopf et al., 1998), respectively. Such lineages have been described in different types of inland waters ranging from lakes and ponds (Jurgens et al., 2000; Auguet & Casamayor, 2008), rivers (Dumestre et al., 2002; Galand et al., 2006) and meromictic lakes (Casamayor et al., 2001; Llirós et al., 2008; Pouliot et al., 2009) and have been recently grouped as the HV-Fresh and PlSA1 lineages (Auguet et al., 2010). LDS and RC-V groups are therefore ubiquitous and very diverse in aquatic ecosystems, suggesting that they have a key functional role, especially in freshwater habitats (Auguet et al., 2010). As no cultured representatives of these groups exist, their physiology and metabolisms remain unknown as well as their role in the ecosystems.

To gain insight into the ecological factors of environmental distribution and phylogenetic structure of such important groups of microbes, we combined both phylogenetic and computational methods with ecological analysis on the large data set of genetic information available for LDS and RC-V groups. We carried out an extensive survey on the public 16S rRNA gene

database and on the sequences from our own monitoring of cold environments (i.e. high mountain lakes and the Arctic) to conduct a detailed community phylogenetic analysis. We also added diversity measurements and macroevolutionary analyses to define and interpret the global ecological relevance of the LDS and RC-V lineages in natural ecosystems. The inference of the rate and patterns of cladogenesis over time and among taxa provided a step forward to more deeply explore the natural history of these microorganisms.

7.2 Methods

7.2.1 Database construction

16S rRNA gene sequences of the *Euryarchaeota* groups Rice Cluster-V (RC-V) and Lake Dagow Sediment (LDS) were obtained from data provided in published papers and from our own work in high-altitude Pyrenean lakes (Auguet & Casamayor, 2008) and Arctic aquatic systems (Galand et al., 2006). The data set was completed with additional sequences directly retrieved from the NCBI nonredundant (nr) database in GenBank by Blast search (searches performed in June 2008, release 166). Only sequences from natural environments were retained, whereas artificial human-generated environments were excluded. The different authors used slightly different primers that covered non-full-overlapping regions of the 16S rRNA gene. This primer effect was, however, minor and less important than the ecosystem effect in the community analyses (Liu et al., 2007). Sequences were screened for quality (no nucleotide ambiguities present), and both sequences shorter than 200 nucleotides length and sequences without basic ancillary environmental data were discarded. The final global data set ended with 482 sequences for RC-V and 368 sequences for LDS.

Additionally, we compiled three basic environmental features from the original publications roughly characterizing the environment from which each sequence was retrieved. We defined those features as habitat (soil, sediment or water), salinity (hypersaline, saline or nonsaline) and oxic status (oxic or anoxic), respectively. Unfortunately, other important environmental variables for *Archaea* such as pH, temperature or trophic state (Auguet et al., 2010) could not be included because of the lack of information in the original publications for the complete set of data or because there was no environmental variation across the selected sequences. All sequences and environmental information used are available as a FASTA file in Data Accessibility (Appendix).

7.2.2 Phylogenetic analyses

Phylogenetic inference was derived after two complementary approaches: First, a *de novo* reconstruction using maximum likelihood (ML) and, second, an insertion method by parsimony into a reference 16S rRNA backbone tree. The ML method is more accurate regarding tree topology and was thus used for the diversity and diversification analyses that rely on the most accurate determination of branch lengths. We also compared topologies by parsimony tree reconstruction and the same conclusions were obtained. The sequences were aligned with MAFFT [v6.603] with the algorithm E-INS-i (Katoh & Toh, 2008). The resultant alignment was manually checked and further trimmed using GBLOCKS to eliminate poorly aligned positions and highly divergent regions making the final alignment, although shorter, more suitable and reproducible for phylogenetic analysis (Castresana, 2000). Thus, the final alignment consisted of 650 characters for LDS and 600 characters for RC-V, respectively. Phylogenies were reconstructed using RAxML [v7.0.3] (Stamatakis, 2006) under the general time reversible (GTR) + GAMMA model with 1,000 bootstrap pseudoreplicates and a ML search of the best tree topology. The 16S rRNA gene of *Nitrosopumilus maritimus* SCM1 was used as outgroup. ML trees are available as Newick files in Data Accessibility (Appendix).

The insertion method, in turn, was used for comparison with other archaeal groups. In this case, the 16S rRNA gene sequences were aligned with the Nearest Alignment Space Termination (NAST) algorithm (DeSantis et al., 2006a) and imported by a parsimony algorithm into the ARB-formatted backbone Greengenes tree (DeSantis et al., 2006b; Ludwig et al., 2004) with a frequency filter to exclude highly variable positions. The archaeal groups *Methanobacteriales* (431 sequences) and *Cenarchaeales* (348 sequences) were selected for comparisons because of (i) their equivalent number of sequences, (ii) similar phylogenetic depth in the general *Archaea* tree and (iii) wide environmental distribution.

7.2.3 Data analysis

We used UniFrac (Lozupone & Knight, 2005) to assess whether evolution had separated the communities that inhabit different environments. The UniFrac test is based on the disparity between the observed β-diversity metric and the metrics of 100 random distributions of the sequences among the environments.

Two phylogenetic measures were used to determine the sequence diversity within each environmental feature and to assess how such genetic diversity

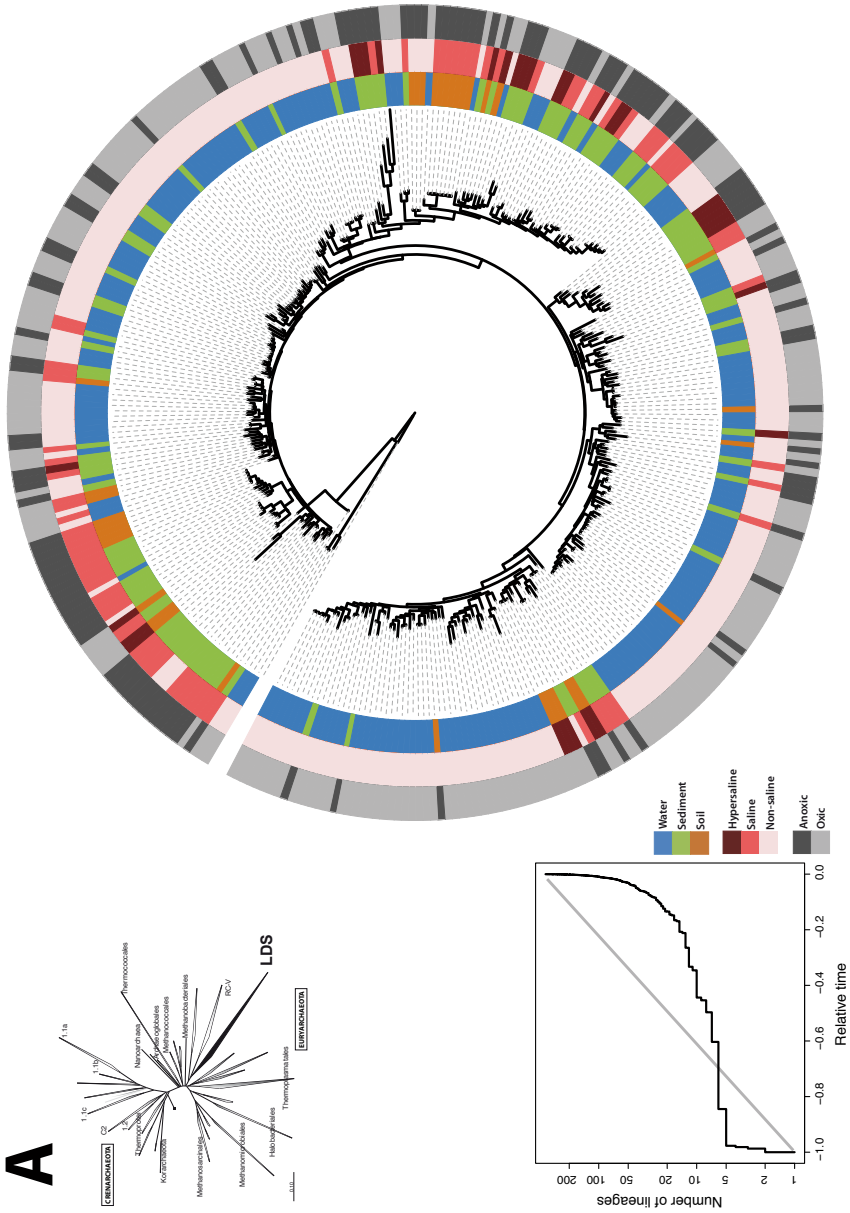
was structured. Phylogenetic diversity (PD) was calculated as the sum of the branch length in a phylogenetic tree (Faith, 1992). To correct for unequal number of sequences, we calculated the mean PD of 1,000 randomized subsamples. The subsample size was the number of sequences present in the smallest data set for each case. The phylogenetic structure was measured with the phylogenetic species variability index (PSV; Helmus et al., 2007). The index quantifies how phylogenetic relatedness declines the variance of a hypothetical neutral trait. The value is 1 when all species are unrelated (i.e. a star phylogeny) and moves towards 0 as species become more phylogenetically related.

We also constructed ultrametric trees (i.e. a tree in which all branch tips are equidistant from the root) by nonparametric rate smoothing (NPRS; Sanderson, 1997) to estimate relative time (scaling to an arbitrary tree age of 1) using molecular phylogenies. Then, we plotted lineage-through-time (ltt) plots and calculated the γ-statistic (Pybus & Harvey, 2000). If diversification has been constant through time, the parameter $\gamma = 0$ and a straight line in the ltt plot is expected. If the diversification slowed, then $\gamma < 0$ and the ltt plot lays above the straight line, while if there is an acceleration in the rate of lineage accumulation, the parameter $\gamma > 0$ and the ltt shows the opposite result (Martin et al., 2004).

All analyses were carried in the R environment (http://www.r-project.org/) using ape (Paradis et al., 2004) and picante (Kembel et al., 2010) packages.

7.3 Results

Figure 7.1 shows the reconstructed ML phylogenetic trees for the euryarchaeal clades LDS (Figure 7.1, panel A) and RC-V (Figure 7.1, panel B) framed within their environmental context. Visually, the clustering already indicated a close relationship between the 16S rRNA gene identities and the environmental features. To deeply explore the phylogenetic ecology of these two clades, we initially tested the association between the environment and the evolutionary process through UniFrac analysis. The results were highly significant (p-value < 0.01) for the environmental features tested (i.e. habitat, salinity and oxic status) and for each euryarchaeal group, suggesting specific evolution within each environment. Most of the sequences were recovered from freshwater ecosystems (approximately 55%), and interestingly, freshwater sequences were basal for each group and closely related to the root of the tree (Figure 7.1).


A
LDS
CRENARCHAEOTA
EURYARCHAEOTA
1.1a
1.1b
1.1c
Q2
Q3
Nanoarchaea
Thermococcales
Korarchaeota
Methanococcales
Methanobacteriales
RC-V
Methanosarcinales
Methanomicrobiales
Halobacteriales
Thermoplasmatales
0.01
Water
Sediment
Soil
Hypersaline
Saline
Non-saline
Anoxic
Oxic
Number of lineages
1
2
5
10
20
50
100
200
Relative time
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

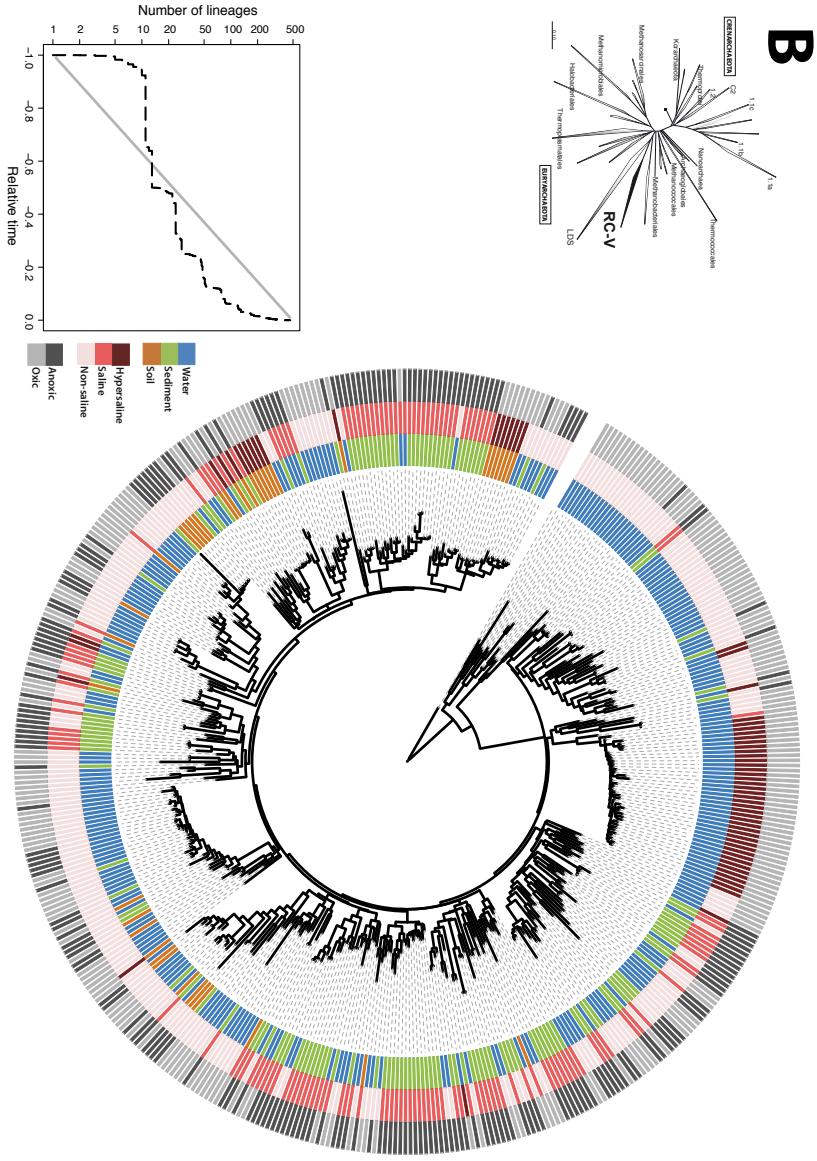


Figure 7.1: *Euryarchaeota* LDS (*panel A*) and RC-V (*panel B*) maximum-likelihood (ML) phylogenetic trees with environmental features mapped as rings. From outer to inner rings: oxic status, salinity and habitat, respectively. Diversification rates plotted as lineage-through-time (ltt) plots based on ultrametric trees (nonparametric rate smoothing, NPRS). The null tree refers to the model of constant diversification through time and $\gamma = 0$. Phylogenetic trees were drawn with iTOL (Letunic & Bork, 2007).

Phylogeny refers to the evolutionary history of a group. To add this view to this ecological study, we plotted the number of lineages in relation to a relative scale of time (ltt plots). In the ltt plots, present time corresponds to relative time 0, whereas relative time -1 is the initial diversification period (Figure 7.1, bottom graphs). The null model tree of expected constant diversification through time (γ = 0) represented by a straight line on a semilogarithmic plot (see in the panel) showed a striking discrepancy to the observed euryarchaeal ltt plots. Both LDS and RC-V showed a consistent increase in the net diversification rate towards present (steeper slope towards relative time 0). Initially, each euryarchaeal group had also rapid lineage accumulation rates that, combined to the recent diversification processes, resulted in high γ parameters (γ = 23.22 for RC-V and γ = 25.10 for LDS).

For both groups, we identified particular diversification events related to the distribution of the environmental features considered. These diversification events may have shaped the observed phylogenetic tree topology of the LDS and RC-V euryarchaeal clades. The LDS phylogeny (Figure 7.1, panel A) indicated that most of the saline and anoxic sequences appeared recently in the history of the group (bootstrap = 100) and then, after the initial expansion, were dominated by freshwater members except for a late diversification event when again new anoxic and saline members emerged. In the case of the RC-V group (Figure 7.1, panel B), most of the recovered sequences from anoxic environments appeared because of a late specific diversification event (bootstrap = 95). Additionally, the RC-V tree clearly illustrated two examples of recently diversified hypersaline and freshwater clusters (bootstrap = 100, in both cases) characterized by a sudden emergence of closely related members.

We also assessed the genetic richness within each environment. Both LDS and RC-V accumulated higher phylogenetic diversity (PD) in oxic freshwaters and sediments than in soils (Table 7.1). Conversely, owing to the evolutionary history mentioned earlier, hypersaline environments were substantially less diverse for RC-V. Hypersaline habitats were also most clustered for RC-V in terms of phylogenetic structure (the lowest PSV value, meaning closely related members and similar lineages; Table 7.1). Therefore, salinity concentration differently shaped the genetic diversity contained within the environments.

Phylogenetic diversity rarefaction curves show how the addition of sequences increases branch length into a given phylogenetic tree (Figure 7.2). In the case of LDS and RC-V, the accumulation of PD did not reach an asymptote evidencing that the current diversity is the tip of the iceberg from the entire diversity existing in the environment. Conversely, a saturated PD rarefaction curve was observed for the archaeal groups used as reference, i.e. *Methanobacteriales* (typical *Euryarchaeota* mostly confined to anoxic marine and terrestrial habitats) and *Cenarchaeales* (ubiquitous marine and freshwater *Crenarchaeota*).

Table 7.1: Phylogenetic measures of the RC-V and LDS inferred trees in relation to all environmental features considered. The mean phylogenetic diversity (PD) of 1,000 randomized subsamples and the phylogenetic species variability index (PSV) are indicated.

	RC-V			LDS		
	Number of sequences	PD	PSV	Number of sequences	PD	PSV
Habitat						
Water	282	7.72	0.50	228	7.67	0.23
Sediment	154	6.57	0.39	106	7.89	0.24
Soil	46	4.86	0.37	34	4.80	0.23
Salinity						
Hypersaline	66	4.14	0.31	30	6.59	0.25
Saline	146	8.17	0.37	78	6.03	0.24
Nonsaline	270	11.26	0.51	260	7.10	0.22
Oxic status						
Anoxic	218	20.47	0.40	130	17.65	0.24
Oxic	264	23.13	0.51	238	20.10	0.24

Thus, the LDS and RC-V groups had much larger levels of genetic diversity when compared with other more deeply studied *Archaea*. In addition, we observed that for the same sampling effort the RC-V group was consistently more phylogenetically diverse than LDS.

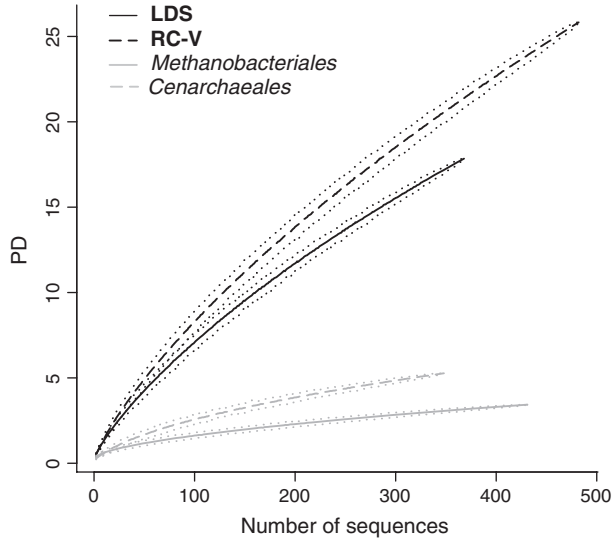


Figure 7.2: Phylogenetic diversity (PD)-based rarefaction curves. The average PD values and standard deviations for 1,000 randomizations are represented.

Finally, as the reconstruction of the microbial phylogenetic trees presented here relied on the 16S rRNA gene, we explored in detail the primary structure of such gene for each *Euryarchaeota* group. We observed that each lineage showed great divergence in the sequence but this low sequence identity was not related to insertions in the highly variable regions of the 16S rRNA gene. This divergence was usually reflected as long-branch lengths in the phylogenetic tree when compared with other archaeal groups. Further, most of the diversity in the LDS and RC-V groups was detected at a 97% sequence identity cut-off (corresponding to the first break in the slope, Figure 7.3). In contrast, the diversity in prokaryotic groups usually relies at a 99% sequence identity level, as illustrated for *Methanobacteriales* and *Cenarchaeales* (Figure 7.3). Altogether, the LDS and RC-V groups contained enormous levels of genetic diversity when compared with other archaeal groups, and proliferation of sequences within each single group was accompanied by significant ecological differentiation.

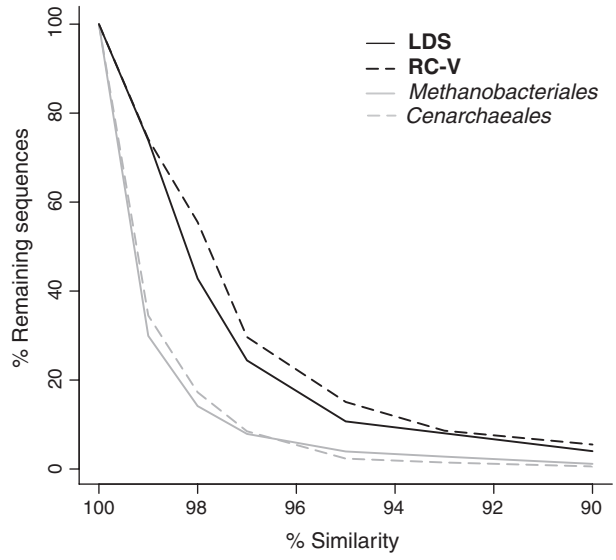


Figure 7.3: Level of diversity accumulation plotted as percentage of remaining sequences versus decreasing values of 16S rRNA gene sequence identity.

7.4 Discussion

Microorganisms are recognized as key players for Earth's ecosystems (Falkowski et al., 2008), but most of the biodiversity and ecology within the microbial world remain poorly known. This holds especially true for the Kingdom *Euryarchaeota* that contains abundant and diverse uncultured inhabitants in aquatic environments with an unknown functional role. In surface marine waters, a few groups of *Euryarchaeota* have the proteorhodopsin gene (Frigaard et al., 2006) that could produce supplemental energy for growth and give a competitive advantage. In deeper water, metagenomic studies have detected Group II *Euryarchaeota* holding putative genes for anaerobic respiratory chains, which might suggest that they gain energy through anaerobic respiration (Martin-Cuadrado et al., 2008). In freshwater ecosystems, however, there is a complete lack of knowledge on the potential physiology and ecology of widespread mesophilic *Euryarchaeota*.

The LDS and RC-V groups have been often detected in nonsaline oxic waters and sediments (Glissman et al., 2004; Galand et al., 2006; Herfort & Jung-Hyun, 2009), and it has been suggested that they are typical freshwater members not adapted to soils (Auguet et al., 2010). Our phylogenetic analysis indicated that freshwater sequences rooted deeply in the LDS and RC-V phylogenetic trees and that they had high genetic diversity, further supporting that they are particularly specific and well adapted to freshwaters. However, it appears that they also had the ability to diversify and successfully colonize other habitats. For instance, we have shown that the RC-V lineage experienced a consistent recent hypersaline transition. Environmental transitions are important processes in the phylogenetic history of *Euryarchaeota* as specific evolution occurs within each habitat (see a recent review by Logares et al., 2009). Although microbial lineages remain specific to macroenvironments over long evolutionary times (Von Mering et al., 2007), the present study shows finer correlations between phylogeny and environment inside narrower phylogenetic clades. However, our observations may be quite conservative because we dealt with general and qualitative environmental features that do not reflect the true complexity of the microhabitats where microorganisms grow.

In general, it has been shown that >50% of microbial 16S rRNA sequences fall into discrete microdiverse clusters containing <1% sequence divergence (Acinas et al., 2004). Such microheterogeneity in the ribosomal sequences has been observed in a wide range of environments and in genetic libraries obtained from *Bacteria*, *Archaea* and eukaryotic microorganisms, suggesting this might be a widespread characteristic of microbial populations (Casamayor et al., 2002). Interestingly, the LDS and RC-V euryarchaeal groups differed from this general observation showing highly diverse phy-

logenetic trees composed of clusters characterized by long-length branches. This characteristic was also reflected in the steepness of the PD rarefaction curves, indicating that there is still larger phylogenetic diversity to be discovered within these *Euryarchaeota*. This adds additional evidence that freshwater ecosystems are good potential targets for the discovery of new archaeal lineages (Auguet et al., 2010). Freshwater habitats are heterogeneous ecosystems with a wide variety of potential functional niches involved in most of the biogeochemical cycles. By contrast, and as in the case of the RC-V group, sequences from hypersaline environments were the less diverse and with a more clustered phylogenetic distribution. This result agrees with the view that extreme environments limit the number of available niches restricting the range of their possible inhabitants (Casamayor et al., 2000a; Estrada et al., 2004).

Remarkably, the LDS and RC-V groups had enormous levels of genetic diversity (particularly the RC-V) when compared with other archaeal groups, raising the question on the cause of this diversity in terms of evolutionary and ecological mechanisms. Phylogenies derived from molecular data provide an indirect record of the speciation events that have led to extant species, reflecting the tempo and mode of macroevolutionary processes related to diversification (Mooers & Heard, 1997). However, the study of macroevolution in microorganisms suffers from two obvious flaws: the lack of fossil records and the unknown scope of the true diversity (Curtis et al., 2002). Different interpretations have arisen when explaining the observed patterns of microbial diversification. Evolutionary considerations stress that different processes can lead to departures from the expectations of the constant speciation rate (Barraclough & Nee, 2001). In a previous meta-analysis using phylogenetic trees inferred for microorganisms, the general pattern conformed to expectations assuming a constant speciation rate ($\gamma = 0$). However, microbial eukaryote trees exhibited more variation than prokaryote trees. The authors hypothesized that the possible effect of horizontal gene transfer (HGT) on prokaryote macroevolutionary dynamics may account for this difference (Martin et al., 2004). Probably in microeukaryotes, successful lateral events tend to occur among more closely related species, or at a lower frequency, than in prokaryotes. Contrastingly, a meta-analysis using 245 molecular phylogenies of macroorganisms (*Arthropoda*, *Chordata*, *Mollusca* and *Magnoliophyta*) indicated that although clades showed great heterogeneity, the average γ was below 0 (McPeek, 2008). Assuming we were working with the most resolute and suitable trees, the observed patterns for the euryarchaeal groups studied here agree with the theory of an actual increase in diversification rate or a constant background extinction rate (high relative to the constant speciation rate). Furthermore, both groups showed rapid lineage accumulation rates in the earliest times but also recent

diversification (γ = 23.22 for RC-V and γ = 25.10 for LDS). Surprisingly, the archaeal groups used as a reference showed the same diversification pattern (γ = 30.46 for *Methanobacteriales* and γ = 20.18 for *Cenarchaeales*), suggesting that similar tree topologies regarding speciation events can be completely different in branch length accumulation. Additionally, these results overlook the possible effect of incomplete sampling because the same pattern was observed for deeply sampled groups (*Methanobacteriales* and *Cenarchaeales*) and groups with a high potential of diversity discovery (LDS and RC-V). In any case, incomplete sampling tends to underestimate the number of nodes towards the present creating an apparent slowdown in diversification rates biasing against phylogenies with $\gamma > 0$ (Barraclough & Nee, 2001).

The missing component in these macroevolutionary models is the ecological context of speciation and extinction. Clades diversify in an ecological context, but most models do not directly encapsulate the ecological mechanisms that influence speciation. A recently published metacommunity model was able to generate the full range of patterns by simply manipulating the degree of ecological differentiation of new species at the time of speciation (McPeek, 2008). According to such results, the initial upturn of lineage accumulation rate that we observed may be attributed to greater ecological opportunities (niche-filling phase). Moreover, the rapid recent diversification could be driven by isolation without ecological differentiation. These accelerating lineage accumulation rates are explained by modes of speciation that generate little or null ecological diversification and therefore are cases where community structure is dominated by neutral ecological drift (McPeek, 2008). The model implies that for a complete understanding of the true macroevolutionary dynamics of ecosystems, the ecological interactions that shaped the history of the clades have to be accounted for. The environments in which freshwater *Euryarchaeota* thrive are remarkably promoting high levels of genetic diversity, although the temporal pattern of diversification (increasing diversification concentrated initially and recently) was the same as in others much less diverse archaeal groups.

Overall, new computational tools permit a more rigorous analysis of the phylogeny of uncultured microorganisms recovered from global environmental surveys (at a broad scale see Auguet et al. 2010; Barberán & Casamayor 2010; for a particular example with freshwater *Actinobacteria*, Newton et al., 2007). As traditionally stated, ecology is the scenario where evolution plays (Margalef, 1963), so it is of great importance to contextualize the phylogenetic information gathered within databases in an ecological context to properly explore the intriguing relationships between environment, diversity and evolution. Here, we have shown the potential of linking phylogeny and environmental description to ascertain the ecological role of abundant but

elusive microbes from natural communities. Species are not independent entities, but their functional and ecological similarities are rather shaped by patterns of common ancestry (Felsenstein, 1985). For uncultured microorganisms (the vast majority of living beings on Earth), how they differ in the genetic or physiological traits used to exploit the environmental resources is largely unknown but we can start exploring ecological causes of speciation patterns through a phylogenetic community ecology framework.

Acknowledgements

This work was supported by grant PIRENA CGL2009-13318-CO2-01/BOS to EOC from the Spanish Ministerio de Ciencia e Innovación (MICINN). PEG was supported by a European Marie Curie grant (CRENARC MEIF-CT-2007-040247) and AFG by CONSOLIDER-INGENIO 2010 project GRACCIE CSD2007-00067 from MICINN. JCA benefits from a Juan de la Cierva-MICINN postdoctoral fellow, and AB is supported by a Spanish FPU predoctoral scholarship. We thank anonymous reviewers for constructive comments.

Appendix

FASTA files (accession numbers and environmental features as header and aligned sequences with gaps) and Newick files of maximum-likelihood inferred trees are available from Dryad: doi:10.5061/dryad.8490.

Part II: Bacterial community ecology in Pyrenean lakes and the influence of Saharan desert dust deposition

8

A phylogenetic perspective on bacterial communities from high-altitude Pyrenean lakes

Resumen

Las regiones montañosas suelen tener un variado mosaico de pequeñas masas de agua que conforman un adiente modelo biogeográfico de ecosistemas cercanos geográficamente pero de gran heterogeneidad ambiental. Además, el carácter aislado y prístino de estos lagos los hace excelentes centinelas y testigos de los cambios climáticos. En este estudio se muestrearon dieciocho lagos pirenaicos basándose en un criterio de selección que maximizase la variación ambiental, y se secuenció el gen ribosomal de la subunidad 16S bacteriana con el objetivo de describir la composición filogenética, su distribución espacial y los patrones de β-diversidad. Los resultados obtenidos demuestran que las comunidades bacterianas presentan una mayor abundancia de *Betaproteobacteria* y una menor abundancia de *Alphaproteobacteria* que otros ecosistemas planctónicos de agua dulce. Se observó que los patrones de β-diversidad estaban estructurados fundamentalmente por el ambiente, siendo el gradiente de pH el más determinante. De manera interesante, se detectó una relación positiva entre el área y la diversidad filogenética, con una pendiente acorde a organismos planctónicos de alta capacidad dispersiva. La aproximación filogenética usada resultó ser una herramienta idónea para los estudios de conservación que recurrentemente se han descuidado de los microorganismos.

Abstract [1]

High mountain lakes districts usually contain a large mosaic of highly diverse small water bodies and thus, they conform a fine biogeographical model of spatially close but environmentally heterogeneous ecosystems. Additionally, their sensitivity to external forcing due to their remoteness and their pristine nature make them excellent sentinels and recorders of environmental changes. We have sampled eighteen high-altitude Pyrenean lakes with a selection criteria focused on capturing the maximum amount of environmental variation within a small geographical area, and sequenced the bacterial 16S rRNA gene in order to describe the phylogenetic community composition, its spatial distribution and β-diversity patterns. Our results showed highly diverse bacterial communities nonrandomly distributed and characterized by higher relative abundance of *Betaproteobacteria* and lower relative abundance of *Alphaproteobacteria* than a global survey of planktonic freshwater ecosystems. Community similarity was primarily structured by the environment, with pH gradient as the strongest driver. Interestingly, we observed a positive relationship between lake area and phylogenetic diversity with a slope consistent with highly dispersive planktonic organisms. We also observed that the bacterial communities from the regional survey in the Pyrenean lakes were more phylogenetically diverse than bacterial communities obtained from coastal marine sites very distant geographically. Our phylogenetic approach incorporated the patterns of common ancestry into community analysis and emerged as a very convenient analytical tool for conservation studies including the recurrently neglected microorganisms.

8.1 Introduction

High mountain lakes districts usually contain a large mosaic of highly diverse small water bodies. This regional diversity driven mostly by the geological characteristics of the catchment contrasts with an overall environmental homogeneity at the global scale (Catalan et al., 2006) and conforms a fine biogeographical model of spatially close but environmentally heterogeneous ecosystems (Reche et al., 2005; Catalan et al., 2009). Additionally, the relative simplicity of high-altitude lakes and their sensitivity to external forcing due to their remoteness from areas of human activity and the pristine nature of their waters make them excellent sentinels and recorders of environmental changes (Catalan et al., 2002). One may expect that diversity within this ecosystem is low regarding that most of the existing high mountain lakes

[1] Barberán A, EO Casamayor. **Manuscript in preparation.**

originated during the last glaciation and thus are young ecosystems, and because high elevation watersheds and mountain ranges restrict colonization from other habitats. In fact, atmospheric deposition is a significant process as a result of their small catchments compared to low land lakes and, for example, represents the main source of nitrogen into the system (Catalan et al., 1994).

Despite the initial expectation, the environmentally heterogeneity of high-altitude lakes at the regional geographical scale, altogether with their pristine and isolated characteristics may enhance the diversity of microorganisms found in this environment (Auguet & Casamayor, 2008; Barberán & Casamayor, 2011). Unfortunately, the microbial composition and its environmental drivers remain poorly explored. Besides their potential for harboring new species, high mountain lakes are good ecosystems for the study of interesting ecological patterns and processes (Prosser et al., 2007). In particular, long-range dispersal of airborne microorganisms (Hervàs & Casamayor, 2009); survival to extreme physical conditions such as low temperatures, low nutrients and high UV exposure (Sommaruga, 2001), and temporal dynamics strictly governed by the ice cover (Felip et al., 1995) may be important in the assembly of local planktonic communities.

To analyze microbial diversity within natural ecosystems and assess its geographical distribution, an approach that takes into account the fact that life forms are not independent entities is essential (Felsenstein, 1985). If ecology is the scenario where evolution plays (Margalef, 1968), ecologists should recognize the historical constraints and incorporate the patterns of common ancestry represented by phylogenies in community analysis (Webb et al., 2002). For this purpose, a set of tools that aim to bridge the gap between evolutionary/phylogenetic analysis and community ecology have been developed recently. Nowadays, ecologists can assess: (i) where most of the biological diversity accumulates (Faith, 1992) and how it is intrinsically structured (Helmus et al., 2007), (ii) how phylogenetic β-diversity (i.e. similarity among communities based on evolutionary history) is distributed along environmental gradients (Bryant et al., 2008; Barberán & Casamayor, 2010), and (iii) how the spatial scale modulates phylogenetic diversity (Morlon et al., 2011).

In this study, we have investigated the spatial distribution of bacterial communities inhabiting high-altitude Pyrenean lakes within a regional scale of a few kilometres using a comprehensive phylogenetic perspective. The central mountain region of the Pyrenees holds the main freshwater lake district of southwest Europe, and constitutes a mosaic of highly diverse water bodies highly coherent with the geological and biological characteristics of the catchments (Catalan et al., 2009). The selected Pyrenean lakes covered a wide range of environmental conditions of more than four pH units, ten-fold phospho-

rous concentration, twenty-fold chlorophyll content, and one order of magnitude of conductivity, being a sample set representative of the landscape heterogeneity found in this lacustrine district. Our specific goals were: (i) to describe the taxonomic composition and the overall phylogenetic diversity and phylogenetic structure of bacterial assemblages in Pyrenean lakes, (ii) to detect the major environmental and geographical factors driving community similarity, and (iii) to extend the traditional framework of island-biogeography theory on species richness including the amount of evolutionary history to its predictions.

8.2 Methods

8.2.1 Study site and sampling

The sampled lakes are located within the Limnological Observatory of the Pyrenees (LOOP; Spanish Pyrenees; 42º33′ N, 00º53′ E) within the protected area of the Aigüestortes National Park (Fig. 8.1). Eighteen lakes were sampled during July 2008 within a two-weeks interval to minimize temporal variability. The lakes ranged between 1,600 and 2,452 m a.s.l., and were selected from different heterogeneous watersheds to cover the regional landscape heterogeneity. The lowest lakes, in the subalpine region below the tree line, were surrounded by meadows. The highest lakes were mostly surrounded by rocky landscape, except Muntanyó d'Àrreu, which was surrounded by montane grasslands and shrublands. The water bodies differed in morphological and limnological characteristics, notoriously in pH, phosphorous and pigments concentrations (Table 8.1). Samples were collected with a Ruttner's sampling bottle on the deepest point of the lakes at a sampling depth 1.5 times the Secchi disk depth, a proxy for the deep chlorophyll maximum. For shallower lakes, samples were collected 1 m above the bottom. The different environmental variables reported in Table 8.1 were measured as recently reported (Mladenov et al., 2011; Auguet et al., In press).

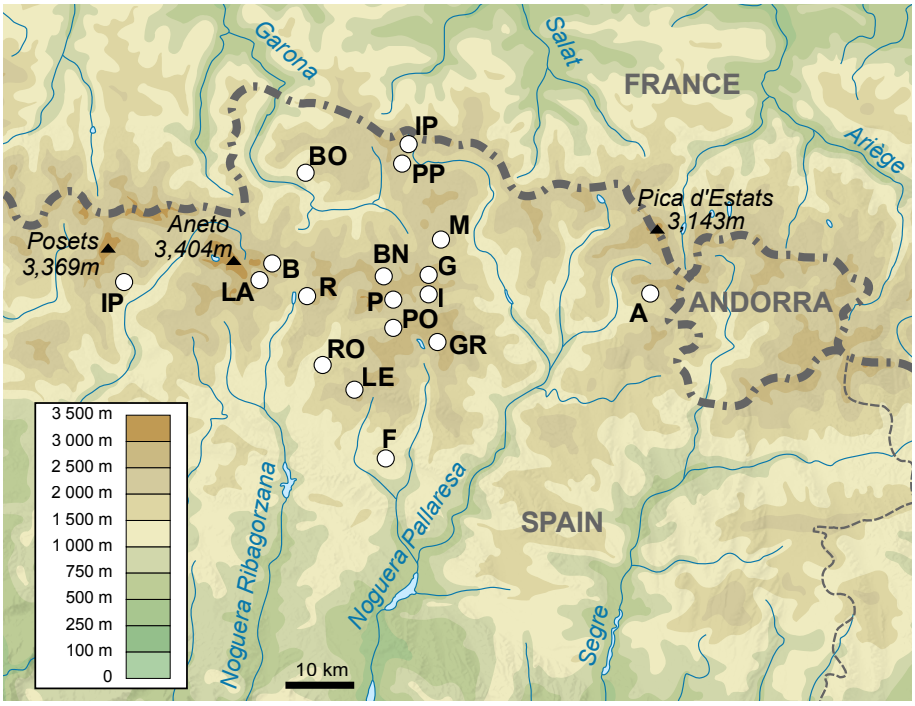


Figure 8.1: Topographic map of the studied area in the mountains of the Pyrenees. National borders, principal rivers and mountain peaks are displayed. White dots indicate the position of the surveyed lakes (see Table 8.1 for abbreviations).

8.2.2 DNA extraction, PCR amplification and cloning

Water samples were prefiltered through a 20-μm net and a 3-μm pore membrane, and then filtered on 0.2-μm pore polycarbonate filters. Polycarbonate filters were incubated with lysozyme, proteinase K, and sodium dodecyl sulfate (SDS) in lysis buffer (40 mM EDTA, 50 mM TRIS, pH 8.3, and 0.75 M sucrose), and phenol extracted as described previously (Dumestre et al., 2002). The bacterial 16S rRNA gene was PCR amplified with primers 27 forward (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492 reverse (5'-GGTTACCTTGTTACGACTT-3') at 48 ºC annealing temperature (Hervàs & Casamayor, 2009). PCR products were purified with the QIAquick PCR Purification kit (Qiagen) and cloned with the TOPO TA cloning kit (Invitrogen) following the manufacturer's instructions. For each library, positive clones were picked and inoculated into LB media in 96-well plates. Clones containing inserts were chosen randomly and were sequenced directly using the

Table 8.1: Geographic coordinates, physico-chemical, morphological and community phylogenetic characterization of the Pyrenean lakes studied. PD: phylogenetic diversity; PSV: phylogenetic species variability; Al/Ac: Lake area/Catchment area.

Lake	Number of sequences	PD	PSV	Latitude (º)	Longitude (º)	Area (ha)	Al/Ac	Altitude (m)	Depth (m)
Aixeus (A)	65	3.06±0.00	0.41	42.61	1.37	3.4	0.042	2,370	15.5
Bassa Nera (BN)	126	4.13±0.34	0.37	42.64	0.92	0.6	0.006	1,890	2
Bassa d'Oles (BO)	106	3.25±0.29	0.25	42.72	0.77	1.3	0.015	1,600	1
Botornat (B)	85	6.71±0.30	0.53	42.59	0.68	3.7	0.013	2,340	22
Filià (F)	153	6.13±0.57	0.52	42.45	0.95	1.4	0.009	2,140	5.5
Gerber (G)	167	5.21±0.47	0.49	42.63	0.99	14.9	0.038	2,170	63
Granotes (GR)	153	4.96±0.42	0.47	42.57	0.97	0.7	0.277	2,330	5
Ibonet Perramó (IP)	81	3.88±0.19	0.44	42.64	0.50	0.7	0.037	2,293	5
L'Illa (I)	131	5.11±0.42	0.44	42.62	0.99	2.1	0.064	2,452	18
Llauset (LA)	88	7.58±0.29	0.54	42.66	0.69	44.3	0.057	2,190	90
Llebreta (LE)	91	5.26±0.32	0.49	42.55	0.89	8.0	0.001	1,620	11.5
Llong de Liat (LL)	138	7.36±0.45	0.53	42.81	0.87	27.1	0.153	2,140	32
Muntanyó d'Àrreu (M)	169	4.30±0.35	0.48	42.67	1.01	4.5	0.041	2,210	14
Pica Palomera (PP)	124	4.77±0.52	0.48	42.79	0.87	4.9	0.072	2,308	10
Plan (P)	142	6.29±0.41	0.51	42.62	0.93	5.0	0.215	2,188	11
Podo (PO)	93	7.59±0.39	0.53	42.60	0.94	4.6	0.139	2,450	20
Redon (R)	95	6.55±0.39	0.48	42.64	0.78	24.0	0.155	2,240	73
Roi (RO)	165	6.57±0.45	0.52	42.57	0.80	3.5	0.030	2,310	10

Lake	Temperature (ºC)	pH	Conductivity (μS·cm^{-1})	DOC (mg·L^{-1})	TDP (μeq·L^{-1})	NH_4^+ (μmol·L^{-1})	NO_3^- (μmol·L^{-1})	Chl a (mg·m^{-3})
Aixeus (A)	8.3	4.97	49.9	0.1	74	0.4	19	0.6
Bassa Nera (BN)	20.1	7.28	53.0	5.5	143	0.6	0	5.3
Bassa d'Oles (BO)	15.2	8.02	154.0	4	294	0.3	0	3.8
Botornat (B)	4.0	7.23	24.1	0.2	168	0.4	14	2.0
Filià (F)	8.7	7.79	133.3	0.3	94	0.6	16	1.7
Gerber (G)	4.9	7.13	23.4	0.6	88	0.3	7	1.3
Granotes (GR)	19.6	6.45	9.8	2.2	119	1	0	4.6
Ibonet Perramó (IP)	13.6	7.49	33.3	0.4	150	0.4	13	0.6
L'Illa (I)	5.7	6.68	13.3	0.4	70	0.7	6	1.3
Llauset (LA)	10.7	7.47	60.4	0.3	113	0.5	16	1.1
Llebreta (LE)	11.3	7.55	30.6	1.1	203	1.1	9	1.8
Llong de Liat (LL)	4.2	7.27	20.8	0.7	68	0.5	2	1.0
Muntanyó d'Àrreu (M)	7.9	9.18	74.3	0.7	674	2.3	0	12.0
Pica Palomera (PP)	11.3	4.61	29.6	0	50	6	4	0.5
Plan (P)	9.2	7.04	16.7	1.4	63	0.4	0	0.6
Podo (PO)	4.7	6.41	9.4	0.3	273	1.2	4	4.2
Redon (R)	5.5	6.60	10.6	0.2	94	1.6	8	0.6
Roi (RO)	8.4	7.06	25.8	0.4	126	1.9	13	5.2

vectors' T7p universal primer. Sequencing reactions were carried out using external facilities (http://www.macrogen.com). Sequences were deposited in GenBank with accession numbers FN296743-FN297704 and HE856822-HE858064, following the MIMARKS (Minimal Information about a marker gene sequence) standards that associates meta-data to each sequence.

8.2.3 Phylogenetic analysis

The 16S rRNA gene sequences were quality checked for nucleotide ambiguities, short sequences (< 400 nucleotides) and possible chimeras using Bellerophon (Huber et al., 2004). The resulting 2,172 sequences were aligned with the NAST alignment tool (DeSantis et al., 2006a) and added by parsimony to the ARB-formatted Greengenes phylogenetic tree after applying a base frequency filter to exclude highly variable positions (Ludwig et al., 2004).

8.2.4 Data analysis

Phylogenetic diversity (PD) for the bacterial community of each lake was calculated as the total sum of branch length (Faith, 1992). To correct for unequal number of sequences, we calculated the mean PD of 1,000 randomized subsamples of each community. The subsample size was the number of sequences present in the smallest dataset. The phylogenetic structure was calculated with the phylogenetic species variability (PSV) index (Helmus et al., 2007). The value is 1 when all species are phylogenetically unrelated (a star phylogeny, totally overdispersed) and approaches 0 as species become more related (clustered). To test whether Pyrenean lakes were composed of bacterial members that are more or less related to each other than expected, we compared the mean observed PSV with the mean of null values using two different randomization procedures. The first null model maintains occurrence (randomizes community data abundances within species), whereas the second null model maintains richness (randomizes community data abundances within samples; Helmus et al., 2007).

We extended the framework of the theory of island biogeography (MacArthur & Wilson, 1967) using PD instead of species richness. A power law scaling between PD and area was assumed

$$PD_{(A)} = cA^z$$

where $\mathrm{PD}_{(A)}$ is the expected phylogenetic diversity contained in a sample from a lake of area A, c is a normalization constant, and z is the slope of the log-transformed function.

Two approaches for ordination by non-metric multidimensional scaling (NMDS) were employed. First, with a Bray-Curtis distance matrix from a table of taxonomic abundances after Hellinger standardization (Legendre & Gallagher, 2001). Second, a distance matrix was constructed using the UniFrac metric, which is a β-diversity metric that quantifies community dissimilarity based on phylogenetic relatedness (Lozupone & Knight, 2005). We used standard and partial Mantel tests to determine the correlation between the UniFrac matrix and geographic (S) and environmental (E) Euclidean distance matrices. Individual environmental predictors were assessed by vector fitting to the NMDS axes and by Permutational Multivariate ANOVA (PERMANOVA; Anderson, 2001). To estimate the degree of spatial autocorrelation of the environmental variables, Moran's coefficient (I) was calculated.

All analyses were carried out in the R statistical environment (http://www.r-project.org) using vegan (Oksanen et al., 2009) and picante packages (Kembel et al., 2010).

8.3 Results

The environmental parameters measured, morphological characteristics and geographic coordinates for the eighteen high-altitude Pyrenean lakes investigated in this study are summarized in Table 8.1. Although distributed over a relatively small area (Fig. 8.1), the lakes covered a wide range of environmental gradients (e.g. lake area from 0.6 to 44.3 ha, from 1 m to 90 m depth, pH from 4.61 to 9.18, or chlorophyll a concentration from 0.6 mg·m^{-3} to 11.99 mg·m^{-3}). Among those variables, only nitrate concentration showed a statistically significant spatial aggregation within our sampling area (Moran's I = 0.10, p-value < 0.05).

The surveyed bacterial communities in the Pyrenean lakes showed high levels of phylogenetic diversity (PD), particularly Podo (PO), Llauset (LA) and Llong de Liat (LL; Table 8.1). As a comparison, we plotted the PD data of a marine survey at nine coastal sites distributed worldwide (Pommier et al., 2007) calculated applying the same sequence processing and rarefaction approach (Fig. 8.2). Bacterial members from high mountain lakes accumulated higher PD per sequence ($2.81{\cdot}10^{-2}$) than marine members ($1.30{\cdot}10^{-2}$). Overall and standardizing by total number of sequences and number of samples because the marine environment was sampled more deeply (c. 336 sequences/sample) than Pyrenean lakes (c. 120 sequences/sample), aquatic bacterial communities from Pyrenean lakes were slightly more phylogenetically diverse ($1.56{\cdot}10^{-3}$) than coastal marine communities very distant geographically ($1.44{\cdot}10^{-3}$).

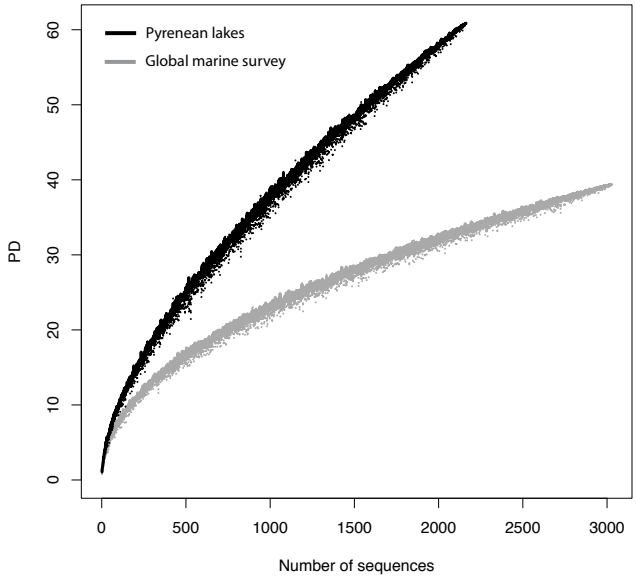


Figure 8.2: Phylogenetic diversity (PD) rarefaction curves. As a comparison, a global survey of marine sites from Pommier et al. (2007) is also plotted.

Bacterial communities from Pyrenean lakes showed a significant phylogenetic structure (see Table 8.1 for details). The mean observed PSV value (0.47) was significantly lower than both the null model maintaining occurrence (0.51, p-value < 0.05) and the null model maintaining richness (0.52, p-value < 0.05) indicating non-random phylogenetic structure. The first null model suggested non-random associations between bacterial sequences among lake communities, with lakes containing more closely related species than expected by chance (phylogenetic clustering), while the second null model suggested that bacterial composition represented non-random samples from the phylogenetic pool (i.e. significant pattern in species prevalence).

We also compared the Pyrenean dataset with data from a previous survey carried out in databases with available 16S rRNA gene sequences from 34 lakes and surface seawaters (Barberán & Casamayor, 2010). At broad taxonomic levels, bacterial composition of Pyrenean lakes was similar to other typologies of lakes distributed worldwide and clearly different from the marine environment (Fig. 8.3). Members of the *Betaproteobacteria*, *Bacteroidetes* and *Actinobacteria* clades dominated the freshwater plankton (for a detailed rarefaction approach of the PD found within each major taxonomic group see the Appendix section). The investigated high-altitude lakes significantly pre-

sented a higher proportion of *Betaproteobacteria* and a lower proportion of *Alphaproteobacteria* than the averaged community present in the global lake survey (p-value < 0.05, t-test; Fig. 8.3).

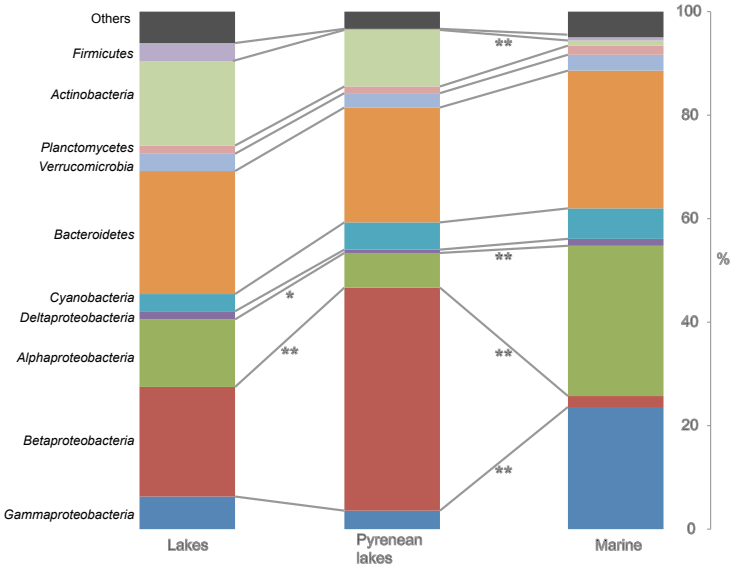


Figure 8.3: Taxonomic composition of the surveyed Pyrenean lakes in this study, and a global comparative analysis of lakes and the marine environment from Barberán & Casamayor (2010). Significant differences are labeled with one asterisk (p-value < 0.05, t-test) or two asterisks (p-value < 0.01, t-test).

Non-metric multidimensional scaling plots of community similarity reflected that bacterial taxonomic and UniFrac distance matrices shared similar patterns as both were strongly correlated (r_M = 0.78, p-value < 0.001, Mantel test; Fig. 8.4). The k-means classification algorithm using the best Calinsky value separated five groups of lakes by their taxonomic composition (see five different colors in Fig. 8.4A). Muntanyó d'Àrreu (M) was the lake with the highest pH (9.18; Table 8.1) and was dominated by *Chlorobi* and *Gammaproteobacteria*. On the other extreme of pH, the acidic lakes Aixeus (A; pH = 4.97) and Pica Palomera (PP; pH = 4.61) showed absence of *Bacteroidetes*, otherwise very abundant in the whole dataset (Fig. 8.3). Additionally, Pica Palomera (PP) and Filia (F) showed a high proportion of *Alphaproteobacteria*. The k-means partition grouped Bassa Nera (BN) and Bassa d'Oles (BO), both shallow and with high DOC concentration ponds, and taxonomically characterized by high relative abundance of *Betaproteobacteria* and low abundance of *Bacteroidetes*. Such shallow ponds also showed the lowest PSV values (that is,

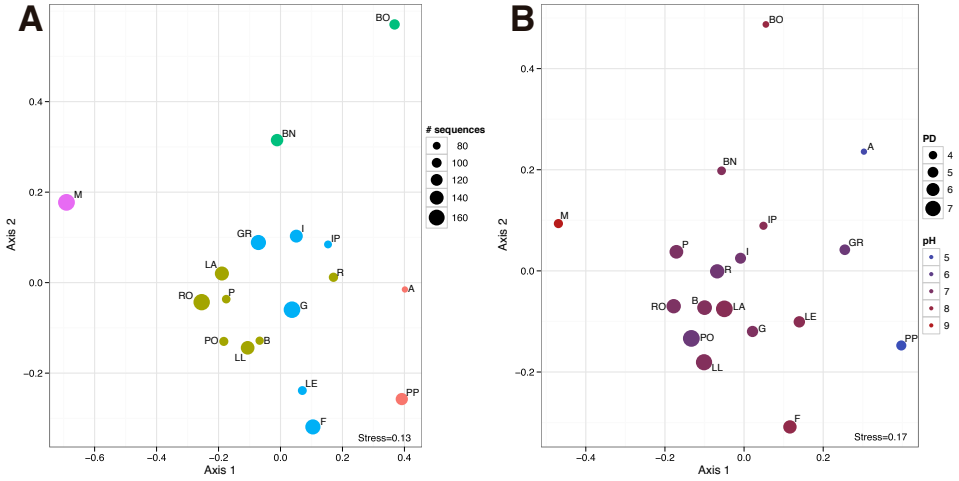


Figure 8.4: Non-metric multidimensional scaling (NMDS) plots of community similarity. (*A*) Based on Bray-Curtis distances of taxonomic abundance after Hellinger standardization. Colours according to a k-means classification at five groups partition (best Calinsky value 5.79) and point size proportional to the number of sequences retrieved from each lake. (*B*) Based on UniFrac distances. Colours according to pH values and point size proportional to phylogenetic diversity (PD).

more phylogenetically clustered bacterial assemblages; Table 8.1). Finally, the k-means classification grouped the remaining lakes in two big heterogeneous groups (Fig. 8.4A).

The phylogenetic β-diversity approach (UniFrac; Fig. 8.4B) confirmed the importance of pH, especially structuring the first axis (R^2 = 0.50, p-value = 0.009, vector fitting) and DOC concentration for the second axis (R^2 = 0.38, p-value = 0.032, vector fitting). Additionally, lakes with the most extreme values of pH had reduced levels of phylogenetic diversity (indicated by point size in Fig. 8.4B). PERMANOVA analyses reinforced the results of vector fitting and selected pH, DOC concentration and conductivity as the most significant variables structuring community similarity patterns (p-value < 0.05). We constructed a Euclidean distance matrix with these three environmental variables standardized (E) and a spatial distance matrix (S) following coordinates to test for biogeographic patterns. Environmental filtering was the most plausible mechanism structuring phylogenetic β-diversity patterns based on Mantel correlations (r_M = 0.65, p-value < 0.001; Fig. 8.5A) and not geographic distance (r_M = 0.10, p-value = 0.307; Fig. 8.5B). The pattern did not change after controlling for possible intermatrix correlations ([E | S]: r_M = 0.65, p-value < 0.001; [S | E]: r_M = 0.08, p-value = 0.359; partial Mantel tests).

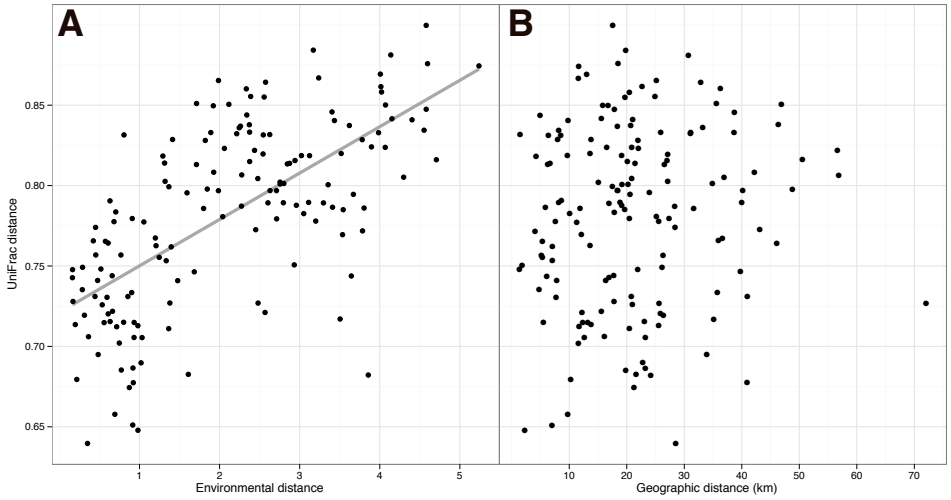


Figure 8.5: Relationship between the UniFrac distance matrix and (*A*) the environmental Euclidean matrix (E) or (*B*) the spatial distance matrix (S). In the first case, a linear regression line is plotted.

Interestingly, we observed a significant correlation between lake area (A) and phylogenetic diversity (PD) of the bacterial assemblages as a whole (r = 0.57, p-value = 0.014). The slope (z) of the log-transformed relationship was 0.13 ± 0.04 (Fig. 8.6). No significant relationships between lake area and PD were found for the two most abundant taxonomic groups, i.e. *Betaproteobacteria* (r = 0.31, p-value = 0.206) and *Bacteroidetes* (r = 0.06, p-value = 0.839; Fig. 8.6). To assess other potential environmental factors influencing PD, we correlated the observed PD with chlorophyll-a concentration (as a surrogate of lake productivity), DOC and TDP (as surrogates of resource richness). No significant results were obtained in any case (Chl a: r = -0.18, p-value = 0.47; DOC: r = -0.42, p-value = 0.08; TDP: r = -0.22, p-value = 0.39).

8.4 Discussion

Although recent sequencing efforts are revealing a large amount of once hidden microbial diversity, some ecosystems still remain largely underexplored such as lakes from high-altitude mountain ranges (Auguet & Casamayor, 2008; Triadó-Margarit & Casamayor, In press). The central mountain range of the Pyrenees constitutes a mosaic of heterogeneous water bodies at the regional scale. These aquatic islands embedded in a sea of land conform an

interesting biogeographical model of spatially close but environmentally distant ecosystem (Catalan et al., 2009). In the present study, we have surveyed eighteen high-altitude lakes from the Pyrenees mountain range with a selection criteria focused on capturing the maximum amount of environmental variation in a small remote area in order to describe the bacterial community composition and their β-diversity patterns.

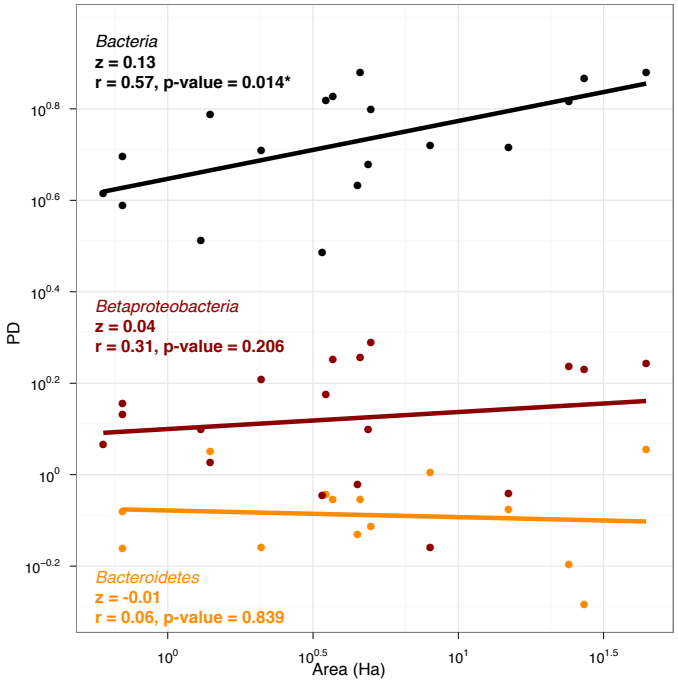


Figure 8.6: Relationship between phylogenetic diversity (PD) and lake area. A linear regression line of the log-transformed relationship is plotted.

The effects of geographic distance and local environmental factors on bacterial community similarity were tested after teasing apart both effects because the spatial dimension may have a spurious correlation with the environmental component. Our results showed that both the overall taxonomic and phylogenetic β-diversity patterns were essentially structured by environmental factors. In particular, pH appeared as the strongest gradient controlling bacterial community patterns. The alkaline (pH= 9.18) lake Muntanyó d'Àrreu (M) showed a high proportion of *Chlorobi* closely related to the anoxygenic phototrophic green sulfur bacterium *Chlorobium phaeobacteroides*, and *Gammaproteobacteria* mainly from the methanotrophic family of the *Methylo-*

coccaceae. These are certainly not alkalophiles but rather specific populations enhanced by the anoxic and sulfurous conditions that develop in the lake hypolimnion. Low pH values, in turn, constrained the presence of the otherwise abundant clade of *Bacteroidetes* in the acidic lakes sampled, Aixeus (A) and Pica Palomera (PP). pH is recognized as one of the major drivers of bacterial community composition (Fierer & Jackson, 2006) and it has been shown to affect the distribution of lineages of freshwater *Actinobacteria* (Newton et al., 2007). In addition, the shallow and eutrophic ponds sampled such as Bassa Nera (BN) and Bassa Oles (BO) appeared distinct due to the low abundance of *Bacteroidetes* and high abundance of *Betaproteobacteria*, most of them closely related to the cosmopolitan freshwater *Polynucleobacter* sp. (Hahn, 2003) and to the methanol-utilizing *Methylophilus* sp. (Salcher et al., 2008). Both groups of *Betaproteobacteria* are particularly abundant in the oxygen-depleted zone of lakes (Salcher et al., 2008). Overall, the bacterial taxonomic composition of Pyrenean lakes was characterized by a higher abundance of *Betaproteobacteria* and a lower abundance of *Alphaproteobacteria* than a global lake survey (Barberán & Casamayor, 2010). *Alphaproteobacteria* were only particularly numerous in lakes Filia (F) and Pica Palomera (PP). The metabolically diverse group of the genus *Rhodobacter* dominated lake Filia (F), while the acidic Pica Palomera lake (PP) was dominated by sequences closely related to the nitrogen-fixing and acid tolerant *Beijerinckia* sp (Barbosa et al., 2002).

Besides shaping the bacterial taxonomic composition and general β-diversity patterns of Pyrenean lakes, environmental conditions may also modulate the amount of phylogenetic diversity present in this environment. The heterogeneous and isolated nature of high-mountain lakes, the influence of the catchment, and the occasional arrival of long-distance immigrants into very diluted waters may promote high phylogenetic diversity. High altitude lakes are probably very efficient collectors of airborne bacteria transported in the atmosphere (Hervàs et al., 2009). In fact, a survey carried out in remote Himalayan lakes located 5,000 m a.s.l. showed abundant populations of bacterial cosmopolitans, closely related to ribosomal sequences retrieved from distant alpine lakes and glaciers (Sommaruga & Casamayor, 2009). This example suggests that the heterogeneous conditions found at the local-regional scale can be minimized when we examine the general pool of globally distributed high mountain lakes. We also noticed that the studied Pyrenean lakes showed more phylogenetic diversity (PD) than that contained after a global ocean survey (Pommier et al., 2007). This observation raises the question of incomplete sampling of microbial communities. It has been shown that sequencing effort does not affect the ranking of richness (Shaw et al., 2008), or β-diversity patterns (Kuczynski et al., 2010). Thus, microbial ecologists should decide whether the efforts for seeking novel microbial life,

with its associated new metabolic potential, focus on a deeper coverage of a particular environment, or to increase the number of samples and to explore isolated and heterogeneous environments and other spatial axes (e.g. switching from the latitudinal or longitudinal axes to the vertical axis; Barberán & Casamayor, 2011).

The species-area relationship is one of the few and oldest general laws in ecology (Arrhenius, 1921; Gleason, 1922). The rich phylogenetic information recovered from ribosomal environmental surveys permits the extension of the traditional predictions of the theory of island biogeography on species richness (MacArthur & Wilson, 1967) in order to incorporate the historical patterns of common ancestry in community ecology analysis (Webb et al., 2002). Along these lines, we observed a positive relationship between lake area and phylogenetic diversity (PD) at the whole community level. This relationship was not significant for the most abundant taxa (*Betaproteobacteria* and *Bacteroidetes*) indicating that bacterial communities conform a coherent unit. Surprisingly, the slope of this relationship was not significantly different from the slope observed using Operational Taxonomic Units (OTUs) as richness estimators instead of PD in a similar aquatic ecosystem (Reche et al., 2005). As discussed in Reche et al. (2005), the accumulation of more biological diversity in larger lakes is more plausible due to higher niche availability rather than to the balance among immigration and extinction as originally proposed by MacArthur & Wilson (1967). The value of the slope is consistent with planktonic organisms with high dispersal rates (Reche et al. 2005; for a recent review see Prosser et al., 2007). The relationship between PD and ecosystem area may not only provide insights into the evolutionary and ecological processes driving community assembly, but also inform conservation policies (Morlon et al., 2011). The PD-area relationship can be used to estimate the potential diversity shrinkage following habitat loss, and how this reduced diversity may affect the response of communities to environmental change (Rodrigues & Gaston, 2002). Incorporating phylogenetic information in conservational ecology is essential because extinctions are not random with respect to phylogeny (Purvis et al., 2000).

In summary, we described the bacterial community composition of a heterogeneous system of Pyrenean high-mountain lakes and found that they are potential targets for the discovery of new microbial diversity. The bacterial PD-area relationship was the expected for planktonic organisms with high dispersion rates. This relationship altogether with the analysis of the environmental constraints that shape β-diversity patterns seem very convenient tools for conservational studies, which have frequently neglected the crucial microbial component.

Acknowledgements

We are thankful to the authorities of the AigüesTortes and Estany de St. Maurici National Park for sampling facilities in the protected areas and continuous support, and to the Centre de Recerca d'Alta Muntanya, Universitat de Barcelona, Vielha for laboratory and logistic facilities. We thank C Gutiérrez-Provecho and Anna Hervàs for laboratory support, and JC Auguet, MA Ballén, L Camarero, M Felip, B Fueyo, JM Gasol, H Sarmento, and M Vila-Costa for fieldwork assistance. We also thank Antoni Fernández-Guerra for computational optimization. This research was supported by grants PIRENA CGL2009-13318 and CONSOLIDER GRACCIE CSD2007-00067 from the Spanish Office of Science and Innovation (MICINN), and AERBAC-2 178/2010 from the Ministerio de Medio Ambiente-Red de Parques Nacionales. AB was supported by the Spanish FPU predoctoral scholarship program.

Appendix

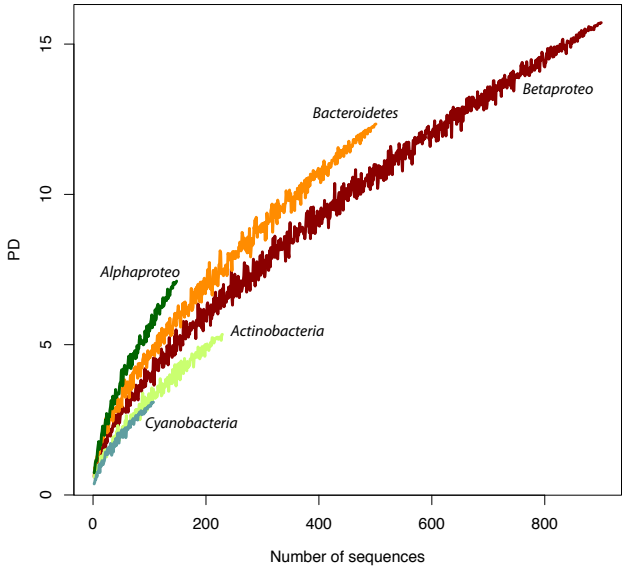


Figure 8.7: Phylogenetic diversity (PD) rarefaction curves for the major taxonomic groups.

9

Structure and temporal patterns in microbial communities transported across continents during Saharan dust events

Resumen

El polvo originado en los grandes desiertos puede viajar largas distancias y ser dispersado sobre superfícies de cientos de miles de kilómetros cuadrados. El proceso de la deposición de polvo está viéndose incrementado a consecuencia del cambio global y constituye una fuente de nutrientes y también un mecanismo para la dispersión microbiana intercontinental. Los lagos alpinos oligotróficos son perfectos centinelas de la magnitud de este proceso, del destino de los microbios colonizadores, y de la calidad microbiológica de la atmósfera global.

En este trabajo, se estudió la composición de las comunidades microbianas de tres hábitats potencialmente conectados atmosféricamente: el origen de las plumas de polvo (i.e. arena del desierto del Sáhara), el transportador atmosférico (i.e. la deposición que llega a la región montañosa de los Pirineos Centrales durante un periodo de tres años) y el colector natural de los

microorganismos aerotransportados (i.e. la interfase aire-agua de los lagos de alta montaña). Se ha descrito la composición taxonómica, los patrones de β-diversidad y la coincidencia en OTUs entre los tres hábitats. El porcentage de OTUs compartidos fue del 10 % del número total de OTUs. Las comunidades microbianas de la deposición atmosférica presentaron un claro patrón temporal. En resumen, el estudio sugiere que existe una variabilidad consistente y predecible en la dinámica y la distribución de las comunidades microbianas aerotransportadas, y que las formas celulares viables con el potencial de colonizar nuevos ambientes no sólo se restringen a las formas esporuladas.

Abstract [1]

Dust originating from the large deserts on Earth, such as the Sahara, can travel long distances and be dispersed over thousands of square kilometres. Dust deposition rates are increasing as a consequence of global change and the dust can represent both a source of nutrients to nutrient-poor ecosystems and a mechanism for intercontinental microbial dispersal. Remote oligotrophic alpine lakes are particularly sensitive to these dust inputs and can serve as sentinels of microbial transport via dust, and of the consequences of the accelerated microbial migration. Here, we studied the microbial community composition of three different airborne dust-connected habitats: the source of dust plumes (i.e. soil from the Saharan desert), the atmospheric carrier (i.e. dust deposited on a high mountain region of the Spanish Pyrenees during a three-year period), and an environment receiving the deposited dust (i.e. the air-water interface of high mountain lakes). We described the taxonomic composition, β-diversity patterns, and OTU overlap across these habitats. The percentage of shared OTUs among the three potentially connected habitats was 10% of the total number of OTUs. The airborne microbial communities of the atmospheric deposition showed a clear temporal pattern with different modules of co-occurring OTUs indicating a consistent and reproducible inter-annual variability. Additionally, we observed the effect of Saharan dust outbreaks. Communities collected less than two days after an outbreak were closer in ordination space to Saharan soil samples than deposition samples collected after periods longer than 40 days without outbreaks. Overall, this study suggests that local and regional features may generate global trends in the dynamics and distribution of airborne microbial assemblages, and that the diversity of viable cells in the high atmosphere that are capable of colonizing new environments is likely higher than previous work has suggested.

9.1 Introduction

Several billion tons (from 0.5 to 5.0) of soil-derived dust are aerosolized each year, transporting soil components between continents at altitudes > 5 km (Kellogg & Griffin, 2006). The Saharan desert in Africa is the largest source of aerosolized soil dust on Earth (50 to 75% of the global dust production), contributing as much as one billion metric tons of dust per year to the atmosphere (Kellogg & Griffin, 2006). The global atmospheric mobilization of dust has likely increased in recent decades due to persistent drought in the Sahara-Sahel region over the past 30 years and an increase in land-use

[1] Barberán A, J Henley, N Fierer, EO Casamayor. **Manuscript in preparation.**

practices (such as livestock grazing) that have desiccated large aquatic regions such as the Chad lake, decreasing vegetation cover and increasing the frequency and intensity of dust storms (Hulme, 2001).

Dust storms can transport microscopic particles thousands of kilometers from their source, a phenomenon that was noted more than 160 years ago by Charles Darwin (Darwin, 1846). Although dust storms may particularly mobilize microscopic organisms such as bacteria and fungi (Bovallius et al., 1978; Prospero et al., 2005); recently, it has been demonstrated that long distance dispersal by wind explains strong floristic affinities among landmasses rather than geographic proximity (Muñoz et al., 2004). This may also be true for microorganisms as we know that bacteria and fungi are commonly found associated with dust (even at altitudes of 20,000m; Griffin, 2004) and many of these microbes can survive prolonged transport in the atmosphere and may even be metabolically active while aloft (Womack et al., 2010). The topic of microbial dispersal via dust events has generated general interest due to concerns about health effects of allergens and the possible long-distance transport of pathogens (Kellogg & Griffin, 2006; Hervàs et al., 2009). In addition, ecologists are interested in understanding the role of these transoceanic and transcontinental dust events in injecting large pulses of microorganisms into the atmosphere, thereby expanding the geographical range of some organisms and possibly altering microbial community composition in depositional environments by facilitating long-distance dispersal events.

High mountain lakes are remote and often pristine ecosystems that are largely unaffected by local anthropogenic factors due to their inaccessibility. For this reason, it has been proposed that high mountain lakes can serve as sentinels of global change (Catalan et al., 2006). Biogeochemistry in these oligotrophic mountain lake environments is dominated by microbial metabolism (Biddanda et al., 2001), and microbial communities have consistent seasonal patterns predictable from snowmelt inputs and lake thermal stability (Pernthaler et al., 1998; Nelson, 2009). In fall, at the Pyrenean region most of the lakes biogeochemistry is driven by dust atmospheric depositions that represent the main input of organic and inorganic compounds to the lakes (Psenner, 1999; Reche et al., 2009; Mladenov et al., 2011). The first natural collector/interceptor of these atmospheric depositions is the neuston, the biological community inhabiting the hydrophobic surface microlayer located within the first millimeter of the air-water interface (Hervàs & Casamayor, 2009). Thus, the neuston of high mountain lakes serves as a useful system for monitoring global-scale microbial dispersal and the effects that long-distance dust deposition may have on microbial communities in aquatic environments.

In this study, we applied high-throughput sequencing techniques (16S rRNA amplicon pyrosequencing) to characterize the microbial communities

living in three different habitats associated with airborne dust: the source of dust plumes (i.e. soil samples from the Saharan desert), the atmospheric carrier (i.e. dust deposition sinking on the Central Pyrenees area fortnightly collected during a period of three years), and the aquatic environment where this dust is first deposited (i.e. the neuston of high mountain Pyrenean lakes). Our objectives were (i) to describe the taxonomic differences and overlap between the microbial communities found in these habitats, (ii) to assess the temporal pattern structuring dust-associated microbial communities, (iii) to explore co-occurrence patterns among microbial immigrants, and (iv) to determine the potential for dust-associated microorganisms to successfully disperse into distant lake environments.

9.2 Methods

9.2.1 Study site and sampling

The sampled lakes are within the Limnological Observatory of the Pyrenees (LOOP; Spanish Pyrenees; 42º33′ N, 00º53′ E) within the protected area of the Aigüestortes National Park (Table 9.1). The neuston (i.e. the air-water surface microlayer) was collected from the upper c. 400 μm of the surface film with a nylon screen sampler as previously reported in Auguet & Casamayor (2008). The lakes were sampled during July 2008 within a two-week interval to minimize temporal variability. Water samples were pre-filtered *in situ* through a 40 μm pore-size net to retain large zooplankton and algae, and 300-500 mL were subsequently filtered on 0.2 μm pore-size polycarbonate filters. The filters were stored in lysis buffer (40 mM EDTA, 50mM Tris pH 8.3, 0.75 M sucrose), enzymatically digested and phenol extracted as reported in Dumestre et al. (2002).

The atmospheric deposition was obtained from an automatic dry/wet passive collector MTX ARS 1010 placed on the vicinity of Lake Llebreta (42.55º N, 0.89º E) at c. 1,800 m altitude during a three-year period approximately twice per month (from 15th May 2007 to 26th May 2010; Table 9.2). Particles deposited in the wet collector (i.e. those washed from the atmosphere by rain or snow precipitation) were collected onto precombusted (450 ºC, 4 hours) Whatman GF/F filters and then, were dried in a laboratory heater for 4 hours and kept in dark (Hervàs et al., 2009). The timing of Saharan dust events was determined by TOMS (Total Ozone Mapping Spectrometer) which provides a measure of the atmospheric loading of UV-absorbing aerosols (i.e. mineral dust and soot from anthropogenic and natural combustion sources; Herman et al., 1997).

Table 9.1: Geographic coordinates, altitude, morphological characteristics, and number of sequences and OTUs at 97% identity of the Pyrenean lakes studied.

Lake	Latitude (º)	Longitude (º)	Altitude (m)	Area (ha)	Depth (m)	Number of sequences	Number of OTUs
Pica Palomera	42.79	0.87	2,308	4.9	10	6,238	407
Certascan	42.70	1.30	2,335	56.9	130	3,579	318
Aixeus	42.61	1.37	2,370	3.4	15.5	4,662	390
Pois	42.65	0.71	2,055	4.4	13	5,056	452
Redon	42.64	0.78	2,240	24	73	2,380	373
Llong Lliat	42.81	0.87	2,140	27.1	32	4,348	536
Granotes	42.57	0.97	2,330	0.7	5	5,247	299
Llebreta	42.55	0.89	1,620	8	11.5	3,135	573
Roi	42.57	0.80	2,310	3.5	10	2,864	390
Bassa Oles	42.72	0.77	1,600	1.3	1	4,818	636
Romedó	42.70	1.32	2,110	11.9	40	4,325	424
Podó	42.60	0.93	2,450	4.6	20	2,577	395
Gerber	42.63	0.99	2,170	14.9	63	4,225	545
Filia	42.45	0.95	2,140	1.4	5.5	5,071	429
Plan	42.62	0.93	2,188	5	11	4,337	520

Table 9.2: Summary of the atmospheric deposition samples analyzed in this study. Mean and standard deviation are indicated for the number of sequences and number of OTUs at 97% identity.

Year	Number of samples	Number of sequences	Number of OTUs
2007	10	4,080±1,435	328±126
2008	14	3,256±1,281	371±209
2009	15	4,194±1,103	334±135
2010	10	3,324±680	361±206

Different samples from Mauritanian sandy soils located within the Sahel region 40 km south-east of Boûmdeid in the Karakoro river basin (c. 3,000 km distant from the Pyrenees) were sampled as source of dust plumes. This area is subjected to frequent dust storms (Kellogg & Griffin, 2006). The soil was treated to obtain soil particles of size < 0.63 mm as reported in Hervàs et al. (2009).

9.2.2 Molecular methods and sequence processing

DNA was extracted using the Mobio PowerSoil DNA Isolation Kit (Mobio Laboratories). Preparation of extracted DNA for pyrosequencing followed the protocol described in detail in Fierer et al. (2008a). In brief, the variable V4 and V5 regions of the 16S rRNA gene (around 250 nucleotides) were am-

plified with the primers F515 (5′-GTGCCAGCMGCCGCGGTAA-3′) and R806 (5′-GGACTACVSGGGTATCTAAT-3′). The F515 primer included the Roche 454-B pyrosequencing adapter and a GT linker, while 806R included the Roche 454-A sequencing adapter, a 12-bp barcode (unique to each sample), and a GG linker. The region amplified by this primer set is well suited for accurate phylogenetic placement of bacterial sequences (Liu et al., 2007) and should amplify nearly all bacteria and archaea with few biases against particular groups (Bates et al., 2010). The resulting barcoded PCR product was normalized in equimolar amounts and sequenced on a Roche GS-FLX 454 automated pyrosequencer at the Environmental Genomics Core Facility (Engencore) at the University of South Carolina.

Raw sequence data generated from pyrosequencing were processed in QIIME (Caporaso et al., 2010). Briefly, sequences were quality trimmed (>200 bp in length, quality score >25, exact match to barcode and primer, and containing no ambiguous characters) and clustered into OTUs with uclust (Edgar, 2010) using both a 97% identity threshold (standard species-level OTU cutoff; Konstantinidis & Tiedje, 2007) and a 90% identity threshold (which corresponds approximately to the taxonomic level of Family for bacteria; Konstantinidis & Tiedje, 2007). We obtained a total of approximately 260,000 sequences from the 68 samples (4 soil, 15 neuston samples and 49 atmospheric deposition samples) with each sample having more than 1,000 sequences. At the 97% identity level, the final OTU table consisted of aproximately 10,500 distinct OTUs, while at the 90% identity the sequences were distributed among 1,700 OTUs. Taxonomic assignment was carried out with the RDP Classifier (Wang et al., 2007), and manually curated by BLAST searches against the GenBank non-redundant nucleotide database (nt).

9.2.3 Data analysis

Community similarity was represented by non-metric multidimensional scaling (NMDS) using the Bray-Curtis distance metric after Hellinger standardization (Legendre & Gallagher, 2001). Analysis of similarities (ANOSIM) was used to test for significant differences between habitat categories. The ANOSIM R statistic is based on the difference of mean dissimilarity ranks between groups and within groups and ranges from 0 (no separation) to 1 (complete separation; Clarke, 1993).

Co-occurrence network inference followed Barberán et al. (2012a). Briefly, we calculated all possible Spearman's rank correlations between OTUs. We considered a valid co-occurrence event to be a robust correlation if the Spearman's correlation coefficient (ρ) was both greater than 0.6 and statistically significant (p-value <0.01; Steinhauser et al., 2008) after adjusting for multiple

testing using the False Discovery Rate (FDR; Benjamini & Hochberg, 1995). In order to describe the topology of the resulting network, a set of measures (i.e., average node connectivity, average path length, diameter, cumulative degree distribution, clustering coefficient, and modularity) were calculated (Newman, 2003). All statistical analyses were carried out in the R environment (www.r-project.org) using vegan (Oksanen et al., 2009) and igraph (Csardi & Nepusz, 2006) packages.

9.3 Results

The three habitats examined here (i.e. soil, deposition, and lake neuston) harbored distinct microbial communities (ANOSIM R = 0.93, p-value < 0.001). The most diverse samples (largest number of different OTUs) were those from Saharan soils. Microbial richness was higher in those deposition samples that were closer in ordination space to soils (Fig. 9.1A). As a whole, neuston and atmospheric deposition bacterial communities resembled more to each other than to the soil assemblage (clustering dendrogram in Fig. 9.1B). Soil samples were more diverse with up to five predominant taxa with relative abundances >10% (*Alpha-* and *Betaproteobacteria*, *Actinobacteria*, *Bacteroidetes*, and *Firmicutes*) whereas only two (*Alpha-* and *Betaproteobacteria*) or four (*Alpha-* and *Betaproteobacteria*, *Actinobacteria*, and *Bacteroidetes*) taxa were found as predominant in deposition and neuston, respectively. *Archaea* were minor components in all the cases (Fig. 9.1B).

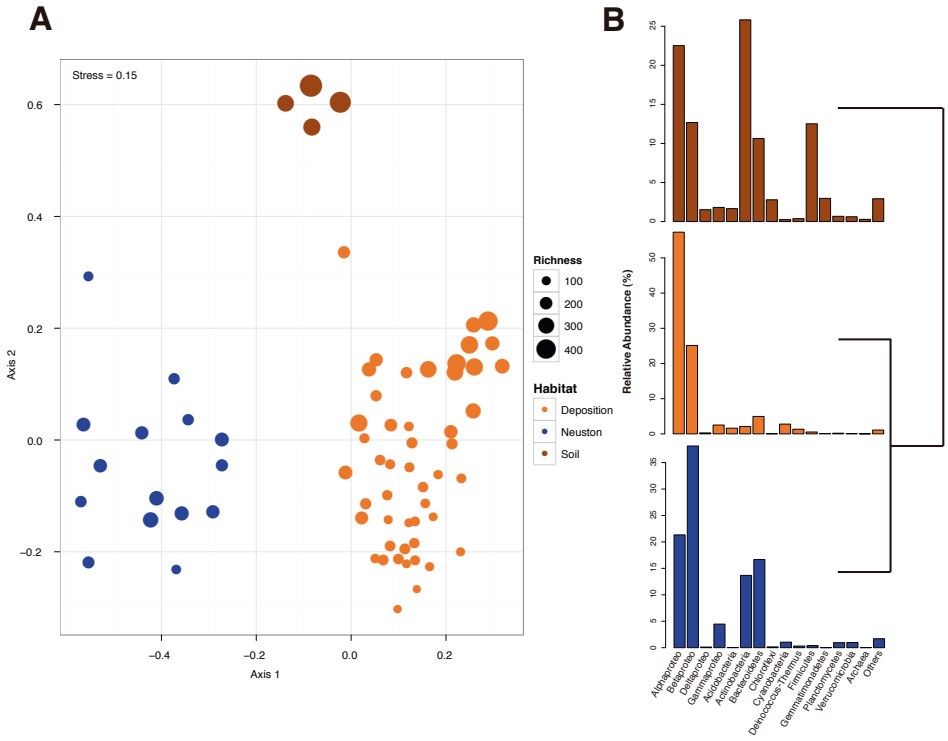


Figure 9.1: (*A*) Non-metric multidimensional scaling (NMDS) ordination plot for the 97% identity OTU table. Size corresponds to richness (number of OTUs). (*B*) Relative abundance of major phylogenetic groups. Habitats were clustered based on average linkage clustering (UPGMA).

Figure 9.2 represents the OTU (at 90% identity) overlap between habitats. Although broad taxonomic composition based on relative abundance showed that atmospheric deposition was more similar to neuston than to soil (see hierarchical clustering in Fig. 9.1B), overall the airborne samples shared more OTUs with soil (321) than with neuston (279). Contrastingly, neuston and soil samples only shared 175 OTUs. The three potentially connected habitats shared a total amount of 154 OTUs (most of them *Alphaproteobacteria* and *Actinobacteria*). Barplots in Fig. 9.2 show the taxonomic composition based on the number of OTUs of the unique habitat compartments and also the pair-wise shared habitat compartments and the total shared. *Alphaproteobacteria*, *Actinobacteria* and *Bacteroidetes* OTUs were the most prevalent in all compartments considered. Besides those groups, soil samples were characterized by the presence of *Chloroflexi* and *Planctomycetes*; deposition by *Firmicutes*, and neuston by *Gammaproteobacteria*. Additionally, some *Betaproteobacteria* and *Firmicutes*

were shared by the three habitats. The pair-wise shared habitat compartments showed an even distribution (specially in shared soil and deposition) with *Betaproteobacteria* more frequent in shared deposition and neuston. As it was the most prevalent group, we examined with more detail the taxonomic composition of the *Alphaproteobacteria* group (Fig. 9.3) looking for differences at the Order level. Interestingly, *Rickettsiales* appeared only in atmospheric deposition samples, while in neuston *Rhodospirillales* predominated and *Rhodobacterales* were less prevalent. Soil samples were more diverse.

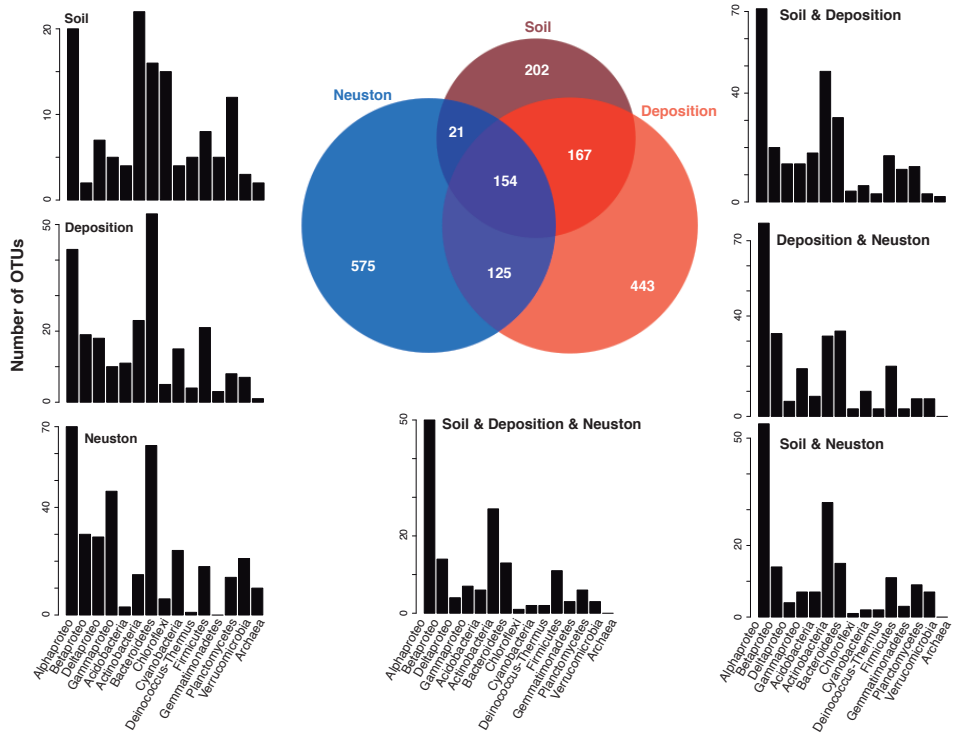


Figure 9.2: Venn diagram showing 90% identity OTU overlap between habitats. Barplots show the taxonomic composition between unique habitats and pair-wise shared habitat compartments.

The temporal trend in Saharan dust outbreaks arriving to the Pyrenees region for the period from May 2007 to May 2010 showed maxima during late spring and summer (from May to September) and a decrease in the number of days per month with presence of airborne Saharan dust (Fig. 9.4A). Interestingly, the temporal decay pattern of community similarity also showed an annual pattern (Fig. 9.4B). During the annual period, deposition samples from disparate dates tended to be more dissimilar than samples closer temporally

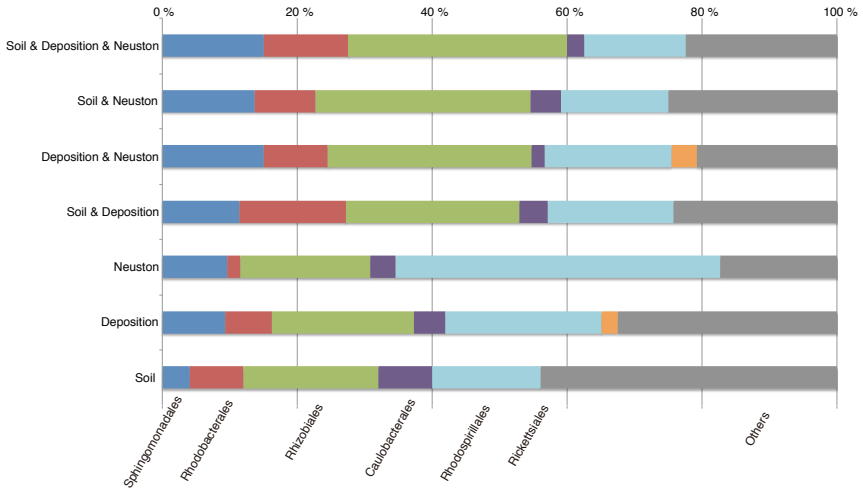


Figure 9.3: Relative abundance of major *Alphaproteobacteria* groups between unique habitats and pair-wise shared habitat compartments.

(linear trend in yellow, Fig. 9.4B). However, this general trend was affected by a strong seasonal pattern (i.e. deposition samples from the same season tend to resemble more each other without regard of the year; local polynomial trend in green, Fig. 9.4B).

Along the survey period different air masses and airborne dusts of local and remote origin can be mixed in the signal explored. To check for bacterial community differences in local vs. remote deposition effect, we split the airborne samples in two contrasted groups: those samples collected less than 2 days after a Saharan deposition event and those with periods longer than 40 days without Saharan dust outbreaks (Fig. 9.5). Interestingly, we noticed significant differences between the groups (ANOSIM R = 0.35, p-value = 0.028). Bacterial communities present in < 2 days samples were closer in ordination space to soil samples than > 40 days samples (Fig. 9.5). The main taxonomic drivers of these differences were *Betaproteobacteria* (higher proportion in < 2 days), *Acidobacteria* and *Verrucomicrobia* (higher proportion in > 40 days; p-value < 0.05, t-test). Miscellaneous unclassified members were also more abundant in > 40 days deposition samples (p-value < 0.05, t-test).

Finally, after assessing general community patterns we applied co-occurrence analysis on the 90% identity OTUs obtained from atmospheric deposition samples to determine common trends in the composition of the airborne bacterial assemblages. We constructed a co-occurrence network based on correlation to estimate how similar were OTU abundances across

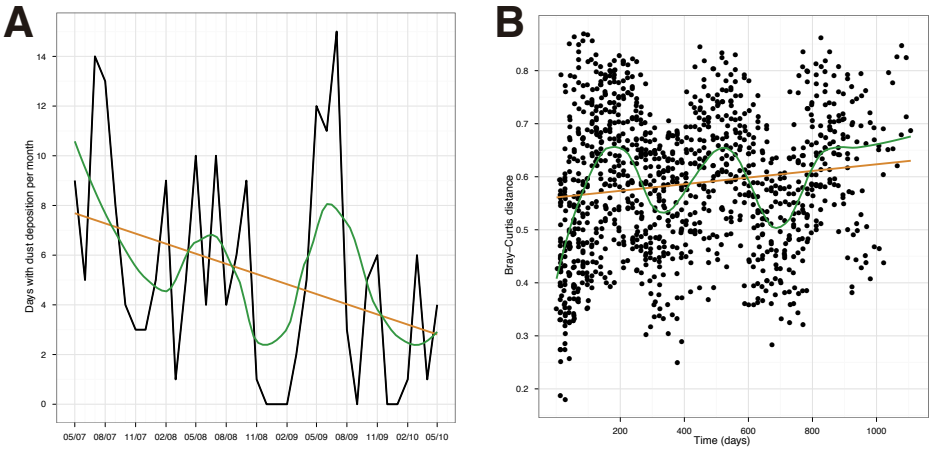


Figure 9.4: (*A*) Days with dust deposition per month during the sampling period. Temporal trend is shown by a fitted linear regression (in yellow) and by a local polynomial regression (loess; in green). Data from www.calima.ws (*B*) Temporal decay pattern of community similarity among atmospheric deposition samples. Trend is shown by a fitted linear regression (in yellow) and by a local polynomial regression (loess; in green).

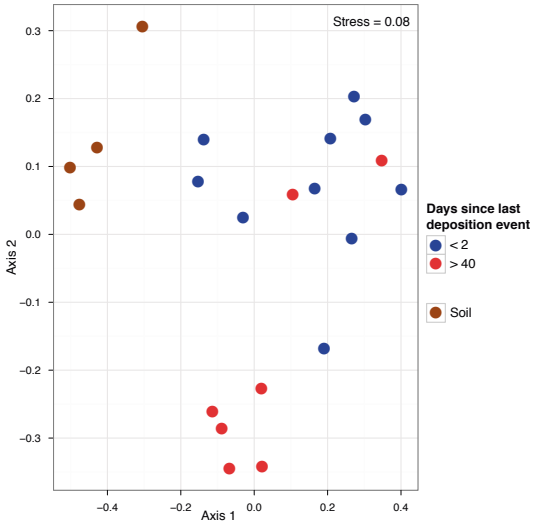


Figure 9.5: Non-metric multidimensional scaling (NMDS) contrasting dust samples in two extreme cases: less than 2 days since last deposition event and more than 40 days since last deposition event. Soil samples are shown as a reference.

the deposition dataset (a link between two nodes in the network correspond to a strong and significant correlation between two OTUs). The resulting microbial network consisted of 186 nodes (OTUs) and 1,761 edges (average degree or node connectivity 18.94; Fig. 9.6A). The topology of the network is described by a few commonly used properties such as the connectivity cumulative degree distribution (Fig. 9.6C). The average network distance between all pairs of nodes (average path length) was 3.04 edges with a diameter (longest distance) of 8 edges. The clustering coefficient (how nodes were embedded in their neighborhood) was 0.52 and the modularity index was 0.30. We further used modularity to detect meaningful modules (i.e. groups of OTUs than tend to co-occur between them more than among the rest). We detected 4 biggest modules that are colored in Fig. 9.6A, and were formed by different taxonomic compositions (Fig. 9.6B). Despite the fact that *Alphaproteobacteria* was the most abundant group in all the modules, we observed additional taxonomic groups that were differently distributed. Thus, *Actinobacteria* and *Bacteroidetes* were very abundant in the blue and green modules. Additionally, *Gemmatimonadetes* and *Acidobacteria* also appeared in the green module. The smallest module (yellow) was also dominated by *Betaproteobacteria*, and *Gammaproteobacteria* of the Order *Pseudomonadales*. The OTUs in the yellow module were abundant in the autumn samples in 2008 and 2009. On the contrary, the OTUs from the blue module were present in winter while the OTUs from the red module were miscellaneously distributed across seasons (though mainly during cold periods) and were associated with most of the samples after long periods (> 40 days) without a Saharan dust outbreak (red points in Fig. 9.3C). OTUs from the green and blue modules were also found in the soil samples, while OTUs from the yellow and blue modules were also present in some of the lake samples. Any module was associated with the summer maxima of Saharan dust outbreaks.

9.4 Discussion

Although the microbial presence in atmospheric aerosols (particularly dust) has been recognized since the 19th century (Darwin, 1846), it has not attracted much multidisciplinary research effort till the last decade specially due to methodological limitations for describing microbial community composition and estimating the magnitude of the atmospheric loads, and because of its recently recognized links with climate change. Generation of atmospheric aerosols and remote dust deposition is a global process, enhanced by perturbations linked to the global change (Moulin & Chiapello, 2006; Prospero & Lamb, 2003). Airborne dust may act both as fertilizer agent and carrier ve-

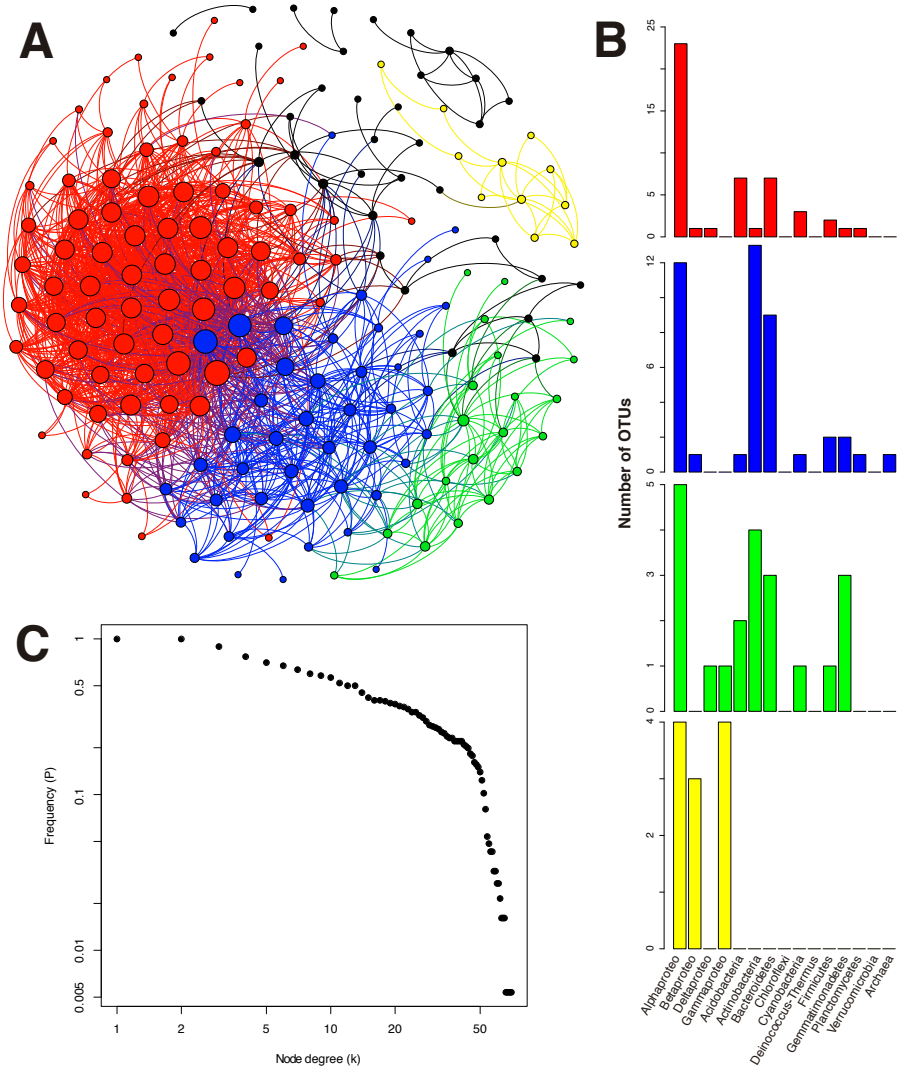


Figure 9.6: (*A*) Network of co-occuring 90% identity OTUs from atmospheric deposition communities. Colours represent modules inside the network (modularity = 0.30). The size of each node is proportional to the number of connections (that is, degree). Network structure was explored and visualized with the interactive platform gephi (Bastian et al., 2009). (*B*) Taxonomic composition of the modules detected. (*C*) Cumulative degree distribution of the network. The connectivity distribution is usually described as a power law distribution. In this case: $P \approx k^{-\alpha}$ with α = 1.30.

hicle for airborne microorganisms. Nutrient inputs from dust depositions increase the bacterial abundance and production in the upper layers of nutrient-limited waters (Herut et al., 2002; Morales-Baquero et al., 2006; Reche et al., 2009; Lekunberri et al., 2010; Mladenov et al., 2011). Furthermore, it is now recognized that dust plumes carry a large set of viable airborne microorganisms with the potential to colonize new environments if the conditions turn favorable (Herut et al., 2002; Hervàs et al., 2009). Despite intercontinental atmospheric dust transport probably plays a key role for global microbial dispersal, little effort has been made to characterize potential routes of immigration for microorganisms.

The atmosphere is not just an intercontinental microbial corridor. Dust particles may provide shadow protection, resources, and a way to capture atmospheric humidity, helping microorganisms to survive the strong desiccation and UV doses during this long journey (Tong & Lighthart, 1998). Our results show that the atmospheric deposition can be considered a truly microbial habitat with a particular community composition, different from potential source and sink habitats. Some deposition-associated microorganisms probably were directly mobilized from the Saharan soils by winds or could have also been injected from local sources (including seawater or vegetation) by advection and vertical mixing in convective clouds. This may explain the high proportion of *Alphaproteobacteria*, a widespread taxon mostly related to seawater (e.g. Barberán & Casamayor, 2010) and soils (e.g. Spain et al., 2009). However, we observed differences in relative abundance at the Order level, e.g. with *Rickettsiales* more associated with atmospheric deposition samples, suggesting that this group has the potential to quickly colonize new remote environments using airborne particles as an effective dispersal vector.

Interestingly, the atmospheric deposition microbial habitat showed a distinguishable annual pattern. Samples temporally apart tend to be dissimilar, and samples from the same season tend to resemble each other suggesting a consistent and reproducible inter-annual variability. Inter-annual recurrence in microbial communities has been also observed in other environments (e.g. Fuhrman et al., 2006; Galand et al., 2010) and apparently it would be also the case here, reinforcing the idea that the atmosphere has environmental characteristics consistent with other habitats where biogeochemical cycling occurs (Womack et al., 2010). It has been proposed that airborne microbial community members may act as ice-nucleation particles (Bowers et al., 2009) and it remains to be tested wheter or not the taxonomic composition of the airborne assemblages may influence cloud dynamics. We also observed both several co-occurrent OTUs associated with different seasons, and a module of OTUs associated with samples from long periods without Saharan intrusions. Recent studies have stressed how the composition of airborne micro-

bial communities is related to land-use type (Bowers et al., 2011a) and season (Bowers et al., 2011b), and probably these local and regional features generate global patterns in the dynamics and distribution of airborne microbial assemblages. In fact, a survey carried out in remote Himalayan lakes located at c. 5,000 m a.s.l., showed most of the bacterial populations detected as cosmopolitan, closely related to ribosomal sequences recovered from distant alpine and glacier regions (Sommaruga & Casamayor, 2009). In addition, a previous study highlighted the short-term temporal variability of airborne bacterial composition with many of the specific taxonomic groups retrieved being common to other cold and highly oligotrophic environments (Fierer et al., 2008b).

Remote lakes are excellent sentinels because they are usually unaffected by direct human influence, yet they receive inputs of atmospheric pollutants, dust, and other aerosols both inorganic and organic. Dust is responsible for the most important difference in precipitation chemistry and, consequently, in the acid-base balance of mountain lakes, and also leads to a significant increase in sulphate, nitrate and ammonium deposition (Psenner, 1999). Furthermore, dissolved organic matter (DOM) from atmospheric deposition may act as an energy source for bacterial growth (Mladenov et al., 2011). Specifically, the deposition maximum from the Sahara to Europe occurs at 2,000-3,000 m altitude (Psenner, 1999) where the lakes sampled in this study were located. Correlations between bacterial abundance, elevation and water residence time suggest that at high elevations, fertilization by dust may be important because of minimal catchment vegetation influences, and relatively greater exposure to atmospheric deposition (Mladenov et al., 2011). *Actinobacteria* which are usually pigmented, spore forming and abundant in high altitude alpine lakes (Warnecke et al., 2005) may succesfully survive this long journey and be potential colonizers of sink aquatic environments. We observed a notable number of actinobacterial OTUs shared by the three potentially connected habitats, suggesting that actinobacterial members might be entering the lakes from the catchment soils as dormant cells or transported for long distances in the troposphere during Saharan dust outbreaks (Hervàs & Casamayor, 2009). Moreover, *Actinobacteria* were abundant in co-occurrent modules (blue and green modules in Fig. 9.6A) associated with soils but also with lake neuston samples. Besides *Alphaproteobacteria* and *Actinobacteria*, OTUs from *Betaproteobacteria*, *Bacteroidetes* and *Firmicutes* were also found in the three distant habitats explored. Recently, some *Gammaproteobacteria* have been proposed as indicators of airborne transport from the Sahara (Hervàs et al., 2009). Overall, it appears that the presence of viable cell forms in the high atmosphere goes far beyond the traditionally considered well-adapted sporulated forms.

In this work we described the community composition from the potential

source (i.e. the Saharan desert), the atmospheric carrier, and the natural collector (i.e. the neuston of high mountain lakes), and the percentage of shared OTUs among the three potentially connected habitats was approximately 10% of the total number of OTUs. This percentage increased to 50% regarding the shared OTU composition of deposition with soil and neuston samples, and up to 88% of the shared soil and neuston OTU composition. Although the impact of microbial atmospheric deposition on sink environments is an important and controversial issue, very few efforts have been addressed to estimate the composition of immigrant microorganisms and the significance of immigration for community assembly (Jones et al., 2008; Hervàs et al., 2009). Colonization via atmospheric deposition may only represent low-level, stochastic variability that is independent of internal dynamics because it has been estimated that deposition over a day only represents 0.0001-0.1% of the bacterial planktonic pool in a lake, and other processes such as particle sinking flux and grazing rates remove equal or greater numbers of cells than estimated to be deposited (Jones et al., 2008). However, the influence on the total lake diversity pool (i.e. the rare biosphere; Pedrós-Alió, 2006) may be paramount due to the extraneous metabolic potential introduced in a pristine environment such as high-altitude lakes. Long-distance dispersal can operate in short-term ecological scale but also as a historical contingency. In our dataset, the detected temporal signal of microbial deposition may influence community assembly in a predictable way. Furthermore, dormancy (which may represent up to 50% of total microbial cell counts in natural environments) may contribute to the potential of long-distance colonization by reducing the risk of local extinction through the recruitment from the diversity pool (Lennon & Jones, 2011).

Overall, our study described the microbial diversity of three different habitats potentially connected by long-distance dispersal. We observed a conspicuous temporal pattern of the microbial communities associated with atmospheric deposition, and co-occurrence patterns among the microbial immigrants. We cannot definitely solve the issue of the significance of microbial colonization via atmospheric deposition, but our results indicate that besides its probable low-level and random variability, the detected temporal signal may point to regular and potentially predictable influences on sink communities. Further modeling (Burrows et al., 2009) and experimental (Hervàs et al., 2009) approaches may enhance our understanding of the connectivity among distant habitats by the intercontinental transport of airborne microorganisms.

Acknowledgements

We are thankful to the authorities of the AigüesTortes and Estany de St Maurici National Park for sampling facilities in the protected areas and continuous support, and to the Centre de Recerca d'Alta Muntanya, Universitat de Barcelona, Vielha for laboratory and logistic facilities. We thank C Gutiérrez-Provecho and X Triadó-Margarit for technical support, JC Auguet, MA Ballén, L Camarero, M Felip, B Fueyo, JM Gasol, H Sarmento, and M Vila-Costa for fieldwork assistance, and José Carlos Brito from CIBIO-Portugal for sand collection in Mauritania. We also thank members of the Fierer lab for helpful discussions, and Anna Hervàs for previous work on the topic and Antoni Fernández-Guerra for QIIME installation and optimization in the CSIC-Blanes server. This research was supported by grants PIRENA CGL2009-13318 and CONSOLIDER GRACCIE CSD2007-00067 from the Spanish Office of Science and Innovation (MICINN), and AERBAC-2 178/2010 from the Ministerio de Medio Ambiente-Red de Parques Nacionales. AB was supported by the Spanish FPU predoctoral scholarship program.

Part III: Novel ecological approaches to pyrosequencing and metagenomics data

10

Using network analysis to explore co-occurrence patterns in soil microbial communities

Resumen

La exploración de extensos conjuntos de datos generados mediante técnicas de secuenciación masiva requiere nuevas aproximaciones analíticas para escudriñar más allá de los inventorios descriptivos. Con el objetivo de investigar potenciales interacciones entre clados de microorganismos, el análisis de redes basados en patrones de co-ocurrencia puede ser una herramienta útil para descifrar la compleja estructura de estas comunidades.

En este trabajo, se aplicó el análisis de redes a un conjunto de datos del gen ribosomal del 16S (más de 160000 secuencias) obtenido por pirosecuenciación de 151 muestras de suelos. Se describió la topología de la red generada y se definieron clados generalistas y especialistas en función de su abundancia y presencia. Gracias a esta nueva perspectiva se pudieron apreciar patrones de co-ocurrencia no aleatorios y estrategias comunes a amplios niveles taxonómicos. En conclusión, se demostró el potencial de la exploración de correlaciones entre clados para la aprehensión de ciertas reglas ecológicas que guían el ensamblaje de las comunidades microbianas.

Abstract [1]

Exploring large environmental datasets generated by high-throughput DNA sequencing technologies requires new analytical approaches to move beyond the basic inventory descriptions of the composition and diversity of natural microbial communities. In order to investigate potential interactions between microbial taxa, network analysis of significant taxon co-occurrence patterns may help to decipher the structure of complex microbial communities across spatial or temporal gradients. Here, we calculated associations between microbial taxa and applied network analysis approaches to a 16S rRNA gene barcoded pyrosequencing dataset containing >160,000 bacterial and archaeal sequences from 151 soil samples from a broad range of ecosystem types. We described the topology of the resulting network and defined operational taxonomic unit categories based on abundance and occupancy (that is, habitat generalists and habitat specialists). Co-occurrence patterns were readily revealed, including general non-random association, common life history strategies at broad taxonomic levels and unexpected relationships between community members. Overall, we demonstrated the potential of exploring inter-taxa correlations to gain a more integrated understanding of microbial community structure and the ecological rules guiding community assembly.

10.1 Introduction

Studies of complex microbial communities have advanced considerably in recent years, in part, due to methodological advances such as high-throughput DNA sequencing technologies that yield detailed information on the composition of microbial communities (Sogin et al., 2006). The sequence data are typically derived from sequencing a portion of the small-subunit rRNA gene (Pace, 1997) and a wide variety of techniques can be applied to the analysis of the sequence data in order to describe the composition of microbial communities, their diversity and how communities may change across space, time, or experimental treatments. However, most of the analytical techniques focus on single properties of the communities being studied. For example, studies describing and comparing the structure of microbial communities often focus on the total numbers of taxa or unique lineages found in individual samples (that is, α-diversity), the relative abundances of individual taxa or lineages and the extent of phylogenetic or taxonomic overlap between communities or community categories (that is, β-diversity). α-diversity measures (for example, richness and coverage estimators, rarefaction curves) yield esti-

[1] See original publication in Barberán et al. (2012a).

mates of microbial diversity and its limits in different environments (Hughes et al., 2001; Curtis et al., 2002; Sogin et al., 2006). Likewise, multivariate statistical techniques such as clustering and ordination have allowed microbial ecologists to describe β-diversity patterns, revealing how biotic and abiotic variables control microbial community composition. For example, analyses of β-diversity patterns have revealed how microbial communities are structured across a wide range of natural habitats (Lozupone & Knight, 2007; Auguet et al., 2010; Barberán & Casamayor, 2010, 2011), the spatial and temporal variability of microbial communities on and in the human body (Fierer et al., 2008a; Costello et al., 2009), and the factors structuring soil bacterial communities (Lauber et al., 2009). In contrast, there has been far less attention focused on using sequence data to explore the direct or indirect interactions between microbial taxa coexisting in environmental samples. Documenting these interactions between taxa (that is, co-occurrence patterns) across complex and diverse communities may help to ascertain the functional roles or environmental niches occupied by uncultured microorganisms (Ruan et al., 2006; Fuhrman & Steele, 2008; Chaffron et al., 2010). With the ever-increasing accumulation of sequence data from microbial communities, we now have the challenge to begin exploring these interactions, and to extend community analyses beyond the exploration of α- and β-diversity patterns that represent the bulk of most sequence-based microbial community analyses.

Network analysis tools and network thinking (Proulx et al., 2005) have been widely used by biologists, mathematicians, social scientists, and computer scientists to explore interactions between entities, whether those entities are individuals in a school (Moody, 2001), species in a food web (Krause et al., 2003), nodes on a computer network (Pastor-Satorras & Vespignani, 2001), or proteins in metabolic pathways (Guimera & Amaral, 2005). Network analysis is used to explore the mathematical, statistical and structural properties of a set of items (nodes) and the connections between them (edges; Newman, 2003). With a few notable exceptions (for example, Ruan et al., 2006; Chaffron et al., 2010; Freilich et al., 2010), network analysis has not been widely applied to exploring co-occurrence patterns between microbial taxa in complex communities. To detect robust associations between microorganisms within and between habitats using network analysis, it is essential to have fairly detailed information on the microbial taxa found across relatively large numbers of samples, as without sufficiently large sample sets it will be difficult to determine whether or not co-occurrence patterns are statistically significant. Ideally, sample sets should cover spatial or temporal gradients in environmental conditions in order for there to be a sufficient variability in taxon abundances to resolve co-occurrence patterns. As shown in recent studies that have used barcoded pyrosequencing to survey microbial communities in large numbers

of samples (for example, Fierer et al., 2008a; Costello et al., 2009; Galand et al., 2009; Lauber et al., 2009), it is now possible to generate microbial datasets that can take full advantage of network analysis approaches and we can apply them to even highly diverse communities, like those found in soils, to explore co-occurrence patterns.

Network analysis of taxon co-occurrence patterns offers new insight into the structure of complex microbial communities, insight that complements and expands on the information provided by the more standard suite of analytical approaches. First, inter-taxa associations may help reveal the niche spaces shared by community members (even members of different domains of life, such as *Bacteria* and *Archaea*) or, perhaps, more direct symbioses between community members. Such information is particularly valuable in environments, such as soil, where the basic ecology and life history strategies of many microbial taxa remain unknown (Janssen, 2006). Exploring co-occurrence patterns between soil microorganisms can help identify potential biotic interactions, habitat affinities, or shared physiologies that could guide more focused studies or experimental settings. More generally, network analysis represents an approach for exploring and identifying patterns in large, complex datasets, patterns that may be more difficult to detect using the standard α/β diversity metrics widely used in microbial ecology (Proulx et al., 2005).

Here we used network analyses to explore associations between prokaryotic taxa in soil, one of the most complex and taxon-rich microbial habitats on Earth. We analyzed over 160,000 bacterial and archaeal 16S rRNA gene sequences from 151 soil samples from a wide variety of ecosystem types in order to demonstrate the utility of network analyses and address the following questions: (i) Do soil microorganisms tend to co-occur more than expected by chance? (ii) Can the lack of agreement between observed and random intra-phyla co-occurrence be used as a proxy of niche differentiation? and (iii) Which taxa are generalists (broadly distributed across soil habitats) or specialists (restricted to certain habitats but locally abundant) and how these ecological categories shape network structure?

10.2 Methods

10.2.1 Soil description and molecular methods

The dataset consisted of 151 soil samples distributed across North and South America, and Antarctica. The collected soils came from a broad range of ecosystems, climates and soil types. Soil collection protocol and methods for edaphic and environmental properties have been described previously (Fierer

& Jackson, 2006; Bates et al., 2010).

Preparation of extracted DNA for pyrosequencing followed the protocol described in detail in Fierer et al. (2008a) and Bates et al. (2010). In brief, a region of the 16S rRNA gene (B250 nucleotides) was amplified with the primers F515 (5′-GTGCCAGCMGCCGCGGTAA-3′) and R806 (5′-GGACTACVSGGGTATCTAAT-3′) that should amplify nearly all bacteria and archaea with few biases against particular groups (Bates et al., 2010). The resulting barcoded PCR product was normalized in equimolar amounts and sequenced on a Roche GS-FLX 454 automated pyrosequencer (Roche Applied Science, Branford, CT, USA) at the Environmental Genomics Core Facility (Engencore) at the University of South Carolina.

10.2.2 Sequence processing

Raw sequence data generated from pyrosequencing were processed in QIIME (Caporaso et al., 2010). Briefly, sequences were quality trimmed and clustered into operational taxonomic units (OTUs) using a 90% identity threshold with uclust (Edgar, 2010). A 90% identity threshold, which corresponds approximately to the taxonomic level of Family for bacteria (Konstantinidis & Tiedje, 2007), was used to generate consistent OTUs with high abundances for subsequent analyses based on correlations and to circumvent potential taxonomic misclassifications due to sequencing anomalies. If we were to use the more standard 'species-level' OTU cutoff (97% sequence identity), the resulting OTU table would be far larger making data visualizations and analysis more difficult. At the 90% identity level, the final OTU table consisted of 160,469 sequences (average of 1,063 sequences per sample) distributed into 4,088 OTUs, of those 2,798 were represented by more than one sequence. Taxonomic assignment was carried out with the RDP Classifier (Wang et al., 2007), and manually curated by BLAST searches against the GenBank nonredundant nucleotide database (nt).

10.2.3 Data analysis

Non-random co-occurrence patterns were tested with the checkerboard score (C-score) under a null model preserving site frequencies (Stone & Roberts, 1990). A checkerboard unit is a 2 x 2 matrix where both OTUs occur once but on different sites. For network inference, we calculated all possible Spearman's rank correlations between OTUs with more than five sequences (1,577 OTUs). This previous filtering step removed poorly represented OTUs and reduced network complexity, facilitating the determination of the core soil community. We considered a valid co-occurrence event to be a robust correlation

if the Spearman's correlation coefficient (ρ) was both > 0.6 and statistically significant (p-value < 0.01, Steinhauser et al., 2008). The nodes in the reconstructed network represent the OTUs at 90% identity, whereas the edges (that is, connections) correspond to a strong and significant correlation between nodes. In order to describe the topology of the resulting network, a set of measures (that is, average node connectivity, average path length, diameter, cumulative degree distribution, clustering coefficient and modularity) were calculated (Newman, 2003). All statistical analyses were carried out in the R environment (http://www.r-project.org) using vegan (Oksanen et al., 2009) and igraph (Csardi & Nepusz, 2006) packages. Networks were explored and visualized with the interactive platform gephi (Bastian et al., 2009).

10.3 Results and discussion

10.3.1 General co-occurrence patterns

Soils are heterogeneous environments that harbor enormously diverse prokaryotic communities (Torsvik et al., 1990; Curtis et al., 2002). Previous work has explored soil microbial diversity from various perspectives, including the estimation of species richness levels in individual samples (for example, Fierer et al., 2007b; Roesch et al., 2007; Youssef & Elshahed, 2008), assessment of the abiotic variables that control the diversity and composition of communities (for example, McCaig et al., 2001; Fierer & Jackson, 2006; Lauber et al., 2009), and the assessment of how specific abiotic factors influence specific taxa (for example, Jones et al., 2009). The relationships between microbial taxa also shape the structure of microbial communities (Prosser et al., 2007), and thus, it can be expected that non-random co-occurrence patterns and significant inter-taxa relationships should occur.

After quality filtering and OTU clustering at the 90% identity, we obtained 2,798 OTUs represented by more than one sequence distributed across the 151 soil samples included in this study. In order to assess non-random co-occurrence patterns, we first used an ecological measure based on checkerboard units (C-score; Stone & Roberts, 1990). Overall, we observed non-random co-occurrence pattern using the whole dataset (C-score = 46.56, p-value < 0.01). Restricting the analysis to only those OTUs showing significant relationships (the ones appearing in Figure 10.1), the measure increased to C-score = 185.03 and p-value < 0.01. A recent meta-analysis showed similar patterns of co-occurrence for microorganisms and macroorganisms suggesting that non-random community assembly may be a general characteristic across all life domains (Horner-Devine et al., 2007). This finding that there are

significant non-random co-occurrence patterns is not surprising, given that we have known for some time that many bacterial taxa exhibit predictable biogeographical patterns (Prosser et al., 2007). Likewise, documenting non-random co-occurrence patterns is far different from actually identifying the causal mechanisms structuring the communities. However, non-random assembly patterns do indicate the dominance of deterministic processes including competitive interactions, non-overlapping niches or historical effects in shaping community composition (Horner-Devine et al., 2007).

10.3.2 Network description

Once we established that the soil microbial assemblage patterns were certainly non-random, we further explored co-occurrence patterns using network inference based on strong and significant correlations (using non-parametric Spearman's; Steinhauser et al., 2008). Correlation networks of co-occurring microorganisms permit the visual summary of lots of information (Chaffron et al., 2010) and have been successfully applied to discern associations between marine microorganisms and their environment (Ruan et al., 2006).

The resulting soil microbial network (Figure 10.1) consisted of 296 nodes (OTUs) and 679 edges (average degree or node connectivity 4.59). Some topological properties commonly used in network analysis were calculated to describe the complex pattern of inter-relationships between OTUs (Newman, 2003). The average network distance between all pairs of nodes (average path length) was 5.53 edges with a diameter (longest distance) of 18 edges. The clustering coefficient (that is, how nodes are embedded in their neighborhood and, thus, the degree to which they tend to cluster together) was 0.33 and the modularity index was 0.77 (values > 0.4 suggest that the network has a modular structure; Newman, 2006). Overall, the soil microbial network was comprised of highly connected OTUs (approximately 5 edges per node) structured among densely connected groups of nodes (that is, modules) and forming a clustered topology (as expected for real-world networks that are more significantly clustered than random graphs). These structural properties offer the potential for quick and easiest comparisons among complex datasets from different ecosystem types in order to explore how the general traits of a certain habitat type may influence the assembly of microbial communities.

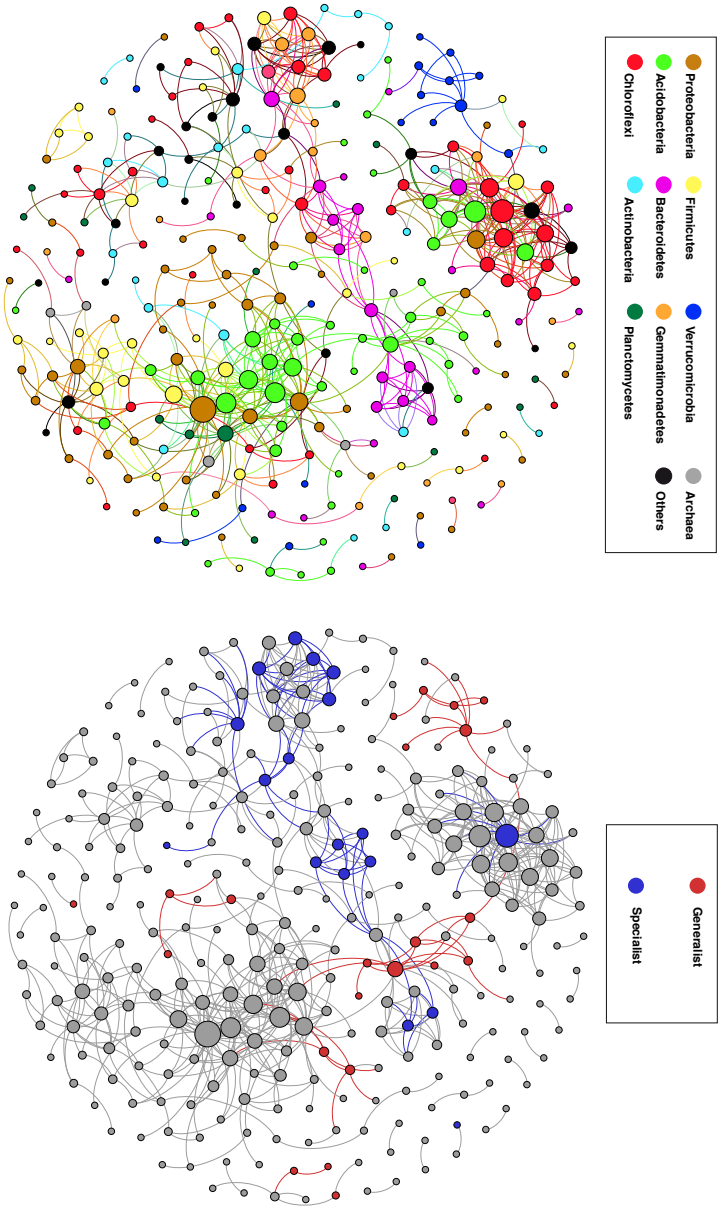


Figure 10.1: Network of co-occurring 90% cut-off OTUs based on correlation analysis. A connection stands for a strong (Spearman's $\rho > 0.6$) and significant (p-value < 0.01) correlation. The size of each node is proportional to the number of connections (that is, degree). *Left panel*: OTUs colored by taxonomy. *Right panel*: OTUs colored by abundance and occupancy (generalists and specialists).

The structural analysis also showed that OTUs from the same phyla tended to co-occur more (33%) than expected when considering observed phyla frequencies and random association (11%; Figure 10.1, left panel). Thus, the magnitude of the lack of agreement between the observed intra-phyla percentage of significant co-occurrences, on the one hand, and the expected assembling under random association, on the other hand, may be used as an index of niche preferences or synergetic relationships. This index may also capture differences among habitats (for example, aquatic vs soil) that may shed light on the ecological rules guiding microbial community composition. In all likelihood, most of these co-occurrence patterns are derived from taxa sharing similar ecological niches, not direct symbioses, however, these data alone do not allow us to separate these two possibilities. Some of the co-occurrence patterns reveal or confirm interesting ecological patterns for taxa that have not been well studied. For instance, members of the verrucomicrobial phylum tended to co-occur more (1.8%) than expected by chance (0.3%) suggesting that, despite being abundant and ubiquitous in soils (Bergmann et al., 2011), they share a specific (and as yet undefined) niche (Figure 10.4). Another abundant and cosmopolitan phylum that also showed higher incidences of co-occurrence than expected by random association was *Acidobacteria* (2.4% at random while 9.4% observed; Figure 10.4). In this case, the pattern is most likely driven by the previously observed phenomenon that soil pH largely governs the distributions of many soil acidobacterial taxa (Jones et al., 2009). Other examples were the *Chloroflexi* and *Deinococcus* groups, which were neither abundant nor generalists/cosmopolitan (Figure 10.4), but instead, appeared to be mostly restricted to desert soils. Several OTUs shared the same habitat preferences and thus appeared to be very interconnected (1.9% at random while 6.2% observed for *Chloroflexi*; 0.01% at random while 1% observed for *Deinococcus*). The degree of disagreement between observed and random co-occurrence may therefore provide further insights in the niche differentiation for the different populations sharing a common phylogeny at different levels of relatedness. Overall, these findings suggest that environmental filtering effects and niche differentiation are evident at broad taxonomic levels, as noted elsewhere (Philippot et al., 2010).

10.3.3 Habitat generalists and specialists

Each of the OTUs represented by more than one sequence was drawn in the abundance vs. occupancy plot (Figure 10.2) to split the set of taxa into two general categories: soil generalists, on the one hand (that is, broadly distributed microbial taxa, which we operationally define here as present in 480 of the 151 soils) and soil specialists (operationally defined here as those that

were locally abundant, representing 42% of the sequences in individual libraries, but only found in < 10 soil samples), on the other hand. Under this criterion, approximately 2% of the OTUs fell into the generalist category while 1% fell into the specialist category. Despite using in the present work a high-throughput DNA sequencing method with higher sequencing depth than traditional methodologies, we may have missed very low abundance taxa that potentially could be habitat generalists. However, dividing taxa from ecological communities into these two categories, which we admittedly defined somewhat arbitrarily, is useful for defining ecological categories/strategies that offer additional information from those defined by phylogeny, taxonomy, or functional capacity (Magurran & Henderson, 2003). Recently, partitioning microbial taxa based on abundance and occupancy has been proved useful in the analysis of clinical samples (van der Gast et al., 2011). In general ecology, positive relationships between mean abundance and occupancy have been observed at many spatial scales (Guo et al., 2000; van der Gast et al., 2011). However, we did not observe such trend in our dataset. Although most of the soil samples analyzed in the present work had their origins in temperate and fertile soils, the environmental variability covered in this study (that is, different habitats and a broad spatial scale) probably altered this relationship. For example, specialist bacterial OTUs inhabiting extreme environments such as deserts or Antarctica soils had a higher abundance than expected regarding their persistence in the overall sampling range.

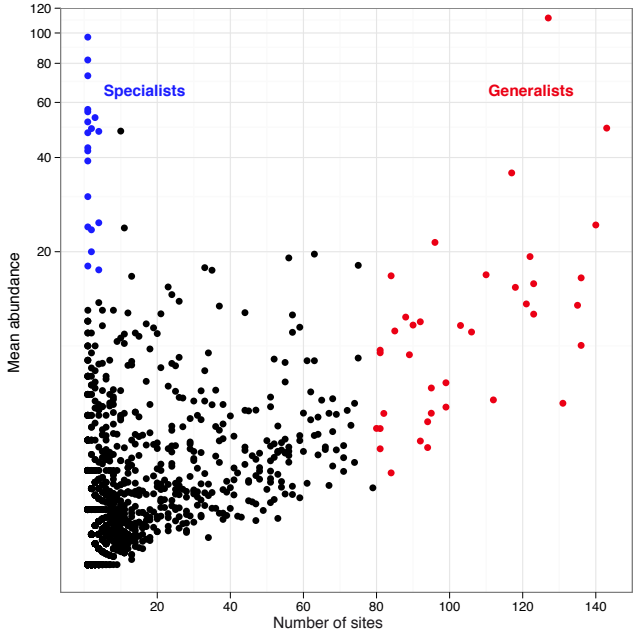


Figure 10.2: Abundance (*y* axis) and occupancy (*x* axis) plot for the 90% cut-off OTUs. Habitat generalists OTUs (in red) defined as appearing in > 80 soil samples. Habitat specialists OTUs (in blue) defined as locally abundant (> 18 sequences) and appearance in < 10 soils.

Overall, we observed a consistent separation in the co-occurring network analysis between generalists and specialists (see significant correlations in Figure 10.1, right panel, against all the remaining OTUs, and a detail in Figure 10.3 adding taxonomic information). Specialists OTUs (locally abundant in a few samples) were composed of a diverse range of phylogenetic groups not common in soils from most biomes (that is, *Chloroflexi*, *Deinococcus*, *Gemmatimonadetes*; Figure 10.3). Generalists OTUs (distributed broadly), in turn, were typical soil members from the *Acidobacteria*, *Proteobacteria* (especially of the *Alpha* subclass) and *Verrucomicrobia* groups (see Figure 10.4, top and low panels, and Janssen 2006 for a recent review). The different taxonomic composition and range of distribution probably influenced the network structure observed (Figure 10.3), indicating that these two ecological categories shaped differently the network structure. The generalists network was less connected and more compartmentalized (19 significant co-occurrences and five compartments) than the specialists network (29 significant co-occurrences and two compartments) probably because of the highest habitat variability covered by the former and the presence in restricted environments by the latter. Thus, the

two somehow arbitrary ecological categories that we established allowed us to capture additional information on the community assembling structure as previously shown for macroorganisms (Pandit et al., 2009).

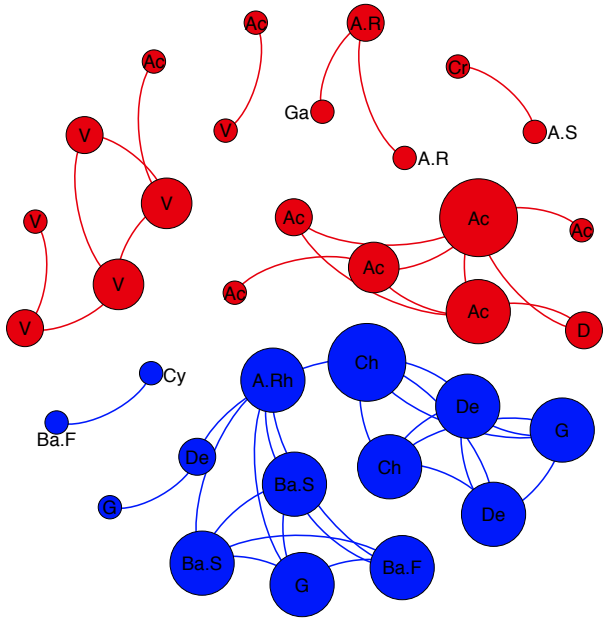


Figure 10.3: Network of co-occurring generalists and specialists 90% cut-off OTUs based on correlation analysis. A connection stands for a strong (Spearman's $\rho > 0.6$) and significant (p-value < 0.01) correlation. The size of each node is proportional to the number of connections (that is, degree). Labels according to taxonomic affiliation: Ac, *Acidobacteria*. A.R, *Alphaproteobacteria*; *Rhizobiales*. A.Rh, *Alphaproteobacteria*; *Rhodobacterales*. A.S, *Alphaproteobacteria*; *Sphingomonadales*. Ba.F, *Bacteroidetes*; *Flavobacteria*. Ba.S, *Bacteroidetes*; *Sphingobacteria*. Ch, *Chloroflexi*. Cr, *Crenarchaeota*. Cy, *Cyanobacteria*. D, *Deltaproteobacteria*. De, *Deinococcus*. G, *Gemmatimonadetes*. Ga, *Gammaproteobacteria*. V, *Verrucomicrobia*.

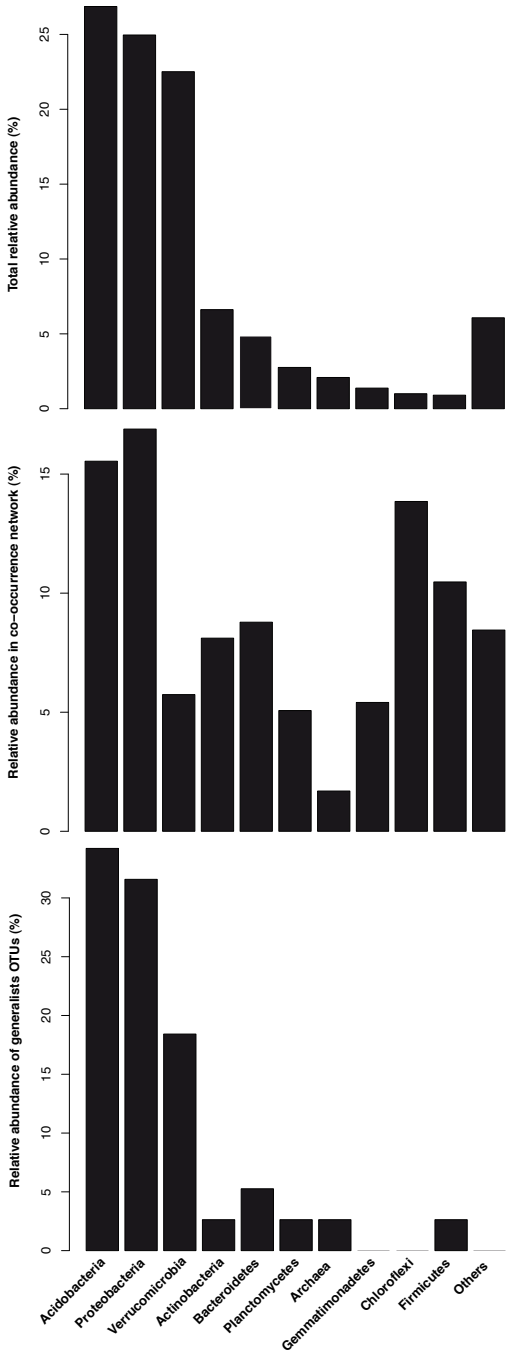


Figure 10.4: Relative abundance of different microbial taxonomic groups. *Top panel*: number of sequences in all soil samples. *Middle panel*: number of significant co-occurrent OTUs (nodes from Figure 10.2). *Low panel*: number of cosmopolitan OTUs.

Finally, the significant correlations found between the generalists OTUs and the whole dataset are shown in Table 10.1. The listed correlations do not include co-occurrent events among members of the same taxonomic group. Generalists OTUs belonged to abundant phyla such as *Verrucomicrobia*, *Acidobacteria*, *Proteobacteria* and *Bacteroidetes*. Interestingly, generalists OTUs not classified as typical and abundant soil members (such as *Deltaproteobacteria* and *Crenarchaeota*) co-occurred with other ubiquitous members. The case of the crenarchaeotal OTU (closely related to Candidatus *Nitrososphaera gargensis* and previously described as ubiquitous in soil; Bates et al., 2010) is particularly remarkable due to our poor understanding of the niches occupied by this taxon, even though it has been proposed that related *Crenarchaeota* may have an important role in the nitrogen cycle as ammonia oxidizers (Leininger et al., 2006). This particular crenarchaeotal OTU co-occurred with sequences closely related to methane oxidizers (Table 10.1). This observation may be a first step to revise the expected functional role of soil *Crenarchaeota* in the nitrification process because of the high sequence similarity of the enzymes that catalyze ammonia oxidation (ammonia monooxygenase) and methane oxidation (particulate methane monooxygenase; Holmes et al., 1995). This is an example of the potential that the approach proposed in the present work has to gain knowledge on elusive but ecologically relevant microorganisms.

10.3.4 Final remarks

With this work we have demonstrated the utility of including network analysis approaches in the repertory of statistical approaches so far available to microbial ecologists. By employing network analyses to a large soil microbial dataset generated by pyrosequencing, the process of exploring the complex set of data was more feasible and interesting unseen patterns emerged, including non-random association, deterministic processes at different taxonomic levels and unexpected relationships between community members. Different ecological rules guiding microbial community composition may be reflected in diverse network structure among habitats (for example, aquatic vs. soil, or early colonized vs. late-successional ecosystems) that deserves further research. The next logical step is to go beyond merely describing the patterns revealed by the network analysis and design more focused experiments, or the study of specific environmental gradients and community shifts over time, in order to understand the mechanisms producing patterns of community coexistence, that is, what finally determines which and how many species live together in a community.

Table 10.1: Taxonomy of generalists OTUs (red nodes in Figure 10.3) and their significant (p-value < 0.01) co-occurrent OTUs. OTUs belonging to *Firmicutes*, *Planctomycetes*, *Actinobacteria*, and *Burkholderiales* and *Rhodocyclales* from the class *Betaproteobacteria* had no significant co-occurrent matches.

Cosmopolitan	Co-occurrent	ρ	Number of co-occurring soils	Closest cultured relative	% identity
Deltaproteo	*Verrucomicrobia*	0.61	53	*Spartobacteria bacterium* Gsoil144	91
	Gammaproteo	0.61	80	*Steroidobacter* sp. ZUMI37	100
	Acidobacteria	0.65	93	*Acidobacteriaceae bacterium* Gsoil1619	90
	Acidobacteria	0.74	102	*Acidobacteria bacterium* IGE-003	93
	Firmicutes	0.61	69	*Moorella thermoacetica*	84
Crenarchaeota	*Alphaproteo; Rhodospirillales*	0.62	64	*Azospirillum amazonense* CBAmc	89
	Alphaproteo; Rhizobiales	0.60	68	*Balneimonas* sp. sptzw09	100
	Bacteroidetes; Sphingobacteria	0.63	82	*Flavosolibacter* sp. HU1-JC5	96
	Alphaproteo; Sphingomonadales	0.71	104	*Sphingomonadaceae bacterium* Gsoil690	100
Verrucomicrobia	*Acidobacteria*	0.71	105	*Acidobacteriaceae bacterium* Gsoil1619	91
	Firmicutes	0.64	97	Unknown sulfate-reducing bacterium clone A8	86
	Acidobacteria	0.68	130	*Bacterium* Ellin6075	97
	Acidobacteria	0.64	70	*Acidobacteria bacterium* IGE-017	91
Acidobacteria	*Verrucomicrobia*	0.68	130	*Verrucomicrobia bacterium* WSF2-44	95
	Actinobacteria	0.62	39	*Rubrobacter xylanophilus* DSM9941	91
	Gammaproteo	0.66	83	*Steroidobacter* sp. ZUMI37	100
	Deltaproteo	0.65	93	*Desulfonatronum* sp. ASO4-1	87
	Crenarchaeota	0.63	78	Candidatus *Nitrososphaera gargensis*	93
	Bacteroidetes; Sphingobacteria	0.63	56	*Sphingobacteriales bacterium* TP524	95
	Gammaproteo	0.62	74	*Gammaproteobacterium* CH43	95
Alphaproteo; Rhizobiales	*Betaproteo; Burkholderiales*	0.61	79	*Burkholderia glathei* YUST-DW12	100
	Gammaproteo	0.64	124	*Proteobacterium* Ellin181	100
Alphaproteo; Sphingomonadales	*Crenarchaeota*	0.71	104	Candidatus *Nitrososphaera gargensis*	93
	Bacteroidetes; Sphingobacteria	0.74	89	*Flavosolibacter* sp. HU1-JC5	96
Gammaproteo	*Deltaproteo*	0.61	33	*Thermodesulforhabdus* sp. nov.M40/2 CIV-3.2	87
	Alphaproteo; Rhizobiales	0.64	124	*Rhodoplanes* sp. 303	97
Bacteroidetes; Sphingobacteria	*Gemmatimonadetes*	0.61	74	*Gemmatimonas aurantiaca* T-27	88

Acknowledgements

We thank members of the Fierer lab, particularly Chris Lauber, Kelly Ramirez and Bob Bowers. We also thank Antoni Fernández-Guerra in the Casamayor lab for QIIME installation and optimization in the CSIC-Blanes server, and members of the Rob Knight lab in UC, particularly Greg Caporaso for computational support and the development of QIIME. AB is supported by the Spanish FPU predoctoral scholarship program and EOC by grants PIRENA CGL2009-13318 and CONSOLIDER-INGENIO 2010 GRACCIE CSD2007-00067 from the Spanish Ministerio de Ciencia e Innovación (MICINN).

11

Exploration of community traits as ecological markers in microbial metagenomes

Resumen

El ritmo de recopilación de información generada por la técnica de la metagenómica está desacoplado de su interpretación ecológica elocuente. Nuevas aproximaciones analíticas basadas en la ecología de caracteres funcionales podrían ser de gran ayuda para solventar este desacoplamiento y extender dicha aproximación al nivel de comunidad en complejos conjuntos de datos genómicos. El objetivo de este estudio fue la exploración de un grupo de caracteres comunitarios que cubrían propiedades tanto nucleotídicas como genómicas en 53 muestras metagenómicas acuáticas provenientes de la expedición GOS.

Como resultado, se encontraron diferencias significativas entre el perfil de β-diversidad derivado del marcador taxonómico del gen ribosomal del 16S y el perfil funcional. Los caracteres analizados discriminaron entre ecosistemas marinos y entre océanos, por lo que se postulan como potenciales marcadores ecológicos. Además, algunas relaciones entre caracteres podrían ser usadas como señales particulares de hábitats o incluso como indicadores de artefactos durante el procesamiento de muestras. Como conclusión, la perspectiva analítica presentada puede ser fructífera para la interpretación de datos metagenómicos dentro de un riguroso marco ecológico.

Abstract [1]

The rate of information collection generated by metagenomics is uncoupled with its meaningful ecological interpretation. New analytical approaches based on functional trait-based ecology may help to bridge this gap and extend the trait approach to the community level in vast and complex environmental genetic data sets. Here, we explored a set of community traits that range from nucleotidic to genomic properties in 53 metagenomic aquatic samples from the Global Ocean Sampling (GOS) expedition. We found significant differences between the community profile derived from the commonly used 16S rRNA gene and from the functional trait set. The traits proved to be valuable ecological markers by discriminating between marine ecosystems (coastal vs. open ocean) and between oceans (Atlantic vs. Indian vs. Pacific). Intertrait relationships were also assessed, and we propose some that could be further used as habitat descriptors or indicators of artefacts during sample processing. Overall, the approach presented here may help to interpret metagenomics data to gain a full understanding of microbial community patterns in a rigorous ecological framework.

11.1 Introduction

In the field of community ecology, there is a resurging interest in understanding biogeographical patterns based on functional traits (i.e. biological characteristics linked to fitness; McGill et al., 2006; Kraft et al., 2008). The study of covarying traits in an environmental context is crucial to understand the ecological strategies that underlie community patterns (Green et al., 2008). In parallel, the new field of metagenomics is challenging the scientific community with an astonishing amount of complex data that intersect the disciplines of microbiology, genetics, ecology and bioinformatics (Handelsman, 2004). Despite some computational advances, the analysis of community genomics data within a meaningful ecological framework remains an elusive goal (Raes et al., 2007a; Kunin et al., 2008). Metagenomics, however, produces data very suitable for extending traditional species-level functional-trait analyses (Wright et al., 2004) to the community level, resulting in an ecological approach to analysing metagenomic data that circumvent the confounding effects of horizontal gene transfer present at lower levels of organization (e.g. at the species or population level).

Recently, the Global Ocean Sampling (GOS) expedition (Rusch et al., 2007) has generated the largest marine metagenomic data set ever sampled along an

[1] See original publication in Barberán et al. (2012b).

environmental gradient, with approximately eight billion nucleotides present in more than 7 million DNA fragments. However, few attempts have been made to analyse this data within an ecological framework (Raes et al., 2011). A synergy between ecology and metagenomics may help bridge this gap, by providing theoretical and analytical tools that could unveil microbial community patterns and the processes that underlie them (Prosser et al., 2007).

To achieve this objective, we have characterized a set of community traits in 53 GOS metagenomic samples taken from the near-surface marine planktonic environment. First, we tested the performance of each trait as a taxonomic, functional and habitat surrogate, respectively. Second, we compared the whole community profile derived from the commonly used 16S rRNA gene marker and from the functional trait set. Finally, we assessed intertrait relationships that could be further used as indicators of functional anomalies and/or for detection of artefacts during sample processing. Overall, the approach presented here is an important step towards developing taxonomic and functional analysis of metagenomic data in a rigorous ecological framework and to provide insights into community ecology beyond purely descriptive studies.

11.2 Methods

11.2.1 Global Ocean Sampling metagenomic data

Unassembled genomic fragments (reads) from the GOS expedition (Rusch et al., 2007) were downloaded from the CAMERA database (Seshadri et al., 2007). We selected 53 surface water samples from picoplankton collected within the same size fraction (0.1–0.8 μm), and free of bacterial contamination during sample handling (DeLong, 2005). Based on current knowledge regarding the spatial and temporal scales of variation in marine microbial communities, a GOS sample represents approximately a week temporally, a few kilometres horizontally and a few metres vertically (Fuhrman, 2009). The analysed metagenomic data set comprised approximately 8,000 Mb contained in approximately 5 million reads.

11.2.2 Community traits calculation

Up to 15 traits were calculated with different level of complexity (see Table 11.1). First, three simple traits were calculated. Custom Perl scripts were used to calculate the GC content and its variance, whereas the odds ratio of dinucleotides was measured as previously described Willner et al. (2009). Din-

ucleotides have been shown to perform better than tri- and tetranucleotides for explanation of habitat differences (Willner et al., 2009).

Next, we extended our approach to three additional traits that relied on the estimation of the number of genomes present in each metagenome. For the assessment of the effective genome size (EGS; Raes et al., 2007b), the number of rRNA genes per genome (Howard et al., 2008) and the number of genes per genome (Biers et al., 2009), we targeted 35 protein markers (see detailed information in Appendix) known to exist as single-copy genes, to be universally distributed along the tree of life, and that are likely recalcitrant to lateral gene transfer (Ciccarelli et al., 2006; Raes et al., 2007b; Wu & Eisen, 2008).

Finally, a set of traits based on the functional annotation of the metagenomic reads was calculated. Automatic annotation of the reads and protein prediction were carried out using MG-RAST (Meyer et al., 2008), which removed strict duplicate reads. Codon and amino acid composition of the predicted proteins were calculated using the program cusp bundled in the emboss package (Rice et al., 2000). The acidic (i.e. glutamic and aspartic acids) to basic (i.e. lysine, histidine and arginine) amino acids ratio (AB) was calculated following (Rhodes et al., 2010). Functional content was based on the comparison against the SEED platform and reported at the subsystem level (Dinsdale et al., 2008). The SEED subsystems are manually curated collections of proteins with related functions (Overbeek et al., 2005). From the functional annotation, we used as traits the percentage of transcriptional factors (TF) and the percentage of SEED subsystems classified reads, both calculated over the reads predicted to be protein coding. Taxonomic content based on 16S rRNA genes was determined by comparing the reads against the Greengenes 16S rRNA gene database and reported at the order level (DeSantis et al., 2006b). For each metagenome, the same parameters were used to ensure the congruity of subsequent analysis. Diversity of the taxonomic and functional contents was calculated using the Shannon index. To correct for unequal sample size, we report the mean of 1,000 randomized subsamples.

Multivariate traits (i.e. taxonomic content, dinucleotides, codons, amino acids and functional content; see Table 11.1) were transformed by considering the projection on the first component of a principal component analysis (PCA).

11.2.3 Statistical analyses

To estimate the degree of spatial autocorrelation of the community traits, Moran's coefficient (I) was calculated. Partial Mantel tests were used to determine the correlation between the similarity of each trait and the taxonomic or functional community similarity. Additionally, analysis of similarities (ANOSIM) was used to test for significant differences within marine ecosys-

tems (coastal vs. open ocean) and between oceans (Atlantic vs. Indian vs. Pacific). The ANOSIM R statistic is based on the difference of mean dissimilarity ranks between groups and within groups and ranges from 0 (no separation) to 1 (complete separation; Clarke, 1993). To test for differences between habitats, PERmutational Multivariate ANOVA (PERMANOVA) was used (McArdle & Anderson, 2001). To represent taxonomic and functional community similarity, we ran nonmetric multidimensional scaling using the Bray–Curtis distance metric after Hellinger standardization (Legendre & Gallagher, 2001). All statistical analyses were carried out in the R environment (http://www.r-project.org) using the ape (Paradis et al., 2004) and vegan (Oksanen et al., 2009) packages.

11.3 Results and discussion

A more complete understanding of microbial processes and patterns is essential to understand ecosystem functions and to predict the Earth's response to global change (Fuhrman, 2009). Community genomics is revealing an unprecedented level of microbial diversity and metabolic novelty in the world's oceans and is the most comprehensive approach currently used to reveal microbial processes and patterns in environmental samples (Handelsman, 2004). To more completely understand these data in an ecological context, we analysed community-level functional traits in 53 selected metagenomic samples from the GOS expedition (Rusch et al., 2007). The analyses produced (i) a defined set of community traits that serve as functional and ecological descriptors of the metagenomic samples; (ii) consistent relationships between traits that may be used for detection of irregularities and/or methodological artefacts and (iii) a different view on microbial communities based either on the taxonomic or on the functional content.

11.3.1 Community traits as functional descriptors of metagenomic samples

We estimated 15 community traits for each metagenomic sample (Table 11.1). To assess the performance of each of the selected traits as a community descriptor, we first tested their spatial autocorrelation. Most of the traits were positively spatially autocorrelated (i.e. closer communities tended to have more similar values), as expected for descriptors of ecological change (i.e. because the environment tends to be spatially autocorrelated). Taxonomic and functional composition showed the highest autocorrelation values (Table 11.1). Although both traits are subject to database scan biases, they sum-

marize two key features of biological communities, that is, community identity and metabolic potential, respectively (Raes et al., 2007a).

The accuracy of the community traits used as descriptors of microbial metagenomes can be potentially related both to a truly functional cause (i.e. different metabolic potentials among different microbial assemblages) or just an effect of community composition (i.e. different taxonomic/phylogenetic groups present in different samples). To distinguish these potential influences, we calculated the correlation between sample similarity for each trait with the taxonomic and the functional composition of each sample (separating the effects of possible intermatrix correlations with partial Mantel tests). Most of the traits showed a significant correlation with functional composition rather than taxonomic composition (Table 11.1), consistent with the hypothesis that they reflect functional differences, rather than just taxonomic differences, among samples. Specifically, GC content, dinucleotides and codon and amino acid compositions (all of them highly correlated) showed the strongest correlation (Table 11.1). Although nucleotidic signatures have been proven useful for the taxonomic assignment of individual genomic fragments (Teeling et al., 2004), some signatures have been successfully applied at the community level for ecological and environmental classification (Willner et al., 2009; Rhodes et al., 2010). For the taxonomic matrix, only taxonomic diversity and the number of genes per genome showed a significant (although weak) correlation (Table 11.1). Overall, the explored community traits were more correlated with the functional composition (r_M = 0.58) than with the taxonomic composition (r_M = 0.42).

We tested the ability of each trait to differentiate between coastal and pelagic communities and among communities from different oceans (Atlantic vs. Indian vs. Pacific). A subset of the community traits (GC content, dinucleotides, codon, amino acids, AB ratio and functional content) was able to clearly distinguish between coastal and open-ocean habitats (Table 11.1). A different subset of traits (number of rRNA per genome, number of genes per genome, taxonomic diversity and taxonomic composition) was effective at distinguishing oceanic origin (Table 11.1). In general, the community traits distinguished slightly better between coastal and pelagic samples (ANOSIM R = 0.33) than the taxonomic composition derived from the commonly used 16S rRNA gene marker (ANOSIM R = 0.31). Nevertheless, taxonomic composition was the best marker of oceanic origin (ANOSIM R = 0.52).

Table 11.1: Summary of the metagenomic community traits explored. Marked with an asterisk, the perentage of variation explained by the first principal component (PCA1) for the traits consisting of a multivariate matrix. Between parentheses, reported p-values. In bold, p-values < 0.01. p-values of partial Mantel correlations (r_M) and ANOSIM R statistic calculated after 999 permutations.

Trait	Mean±sd / PCA1*	Autocorrelation (Moran's I)	Tax | func (r_M)	Func | tax (r_M)	Coastal vs. Open ocean (ANOSIM R)	Atlantic vs. Indian vs. Pacific (ANOSIM R)
GC content	37.7±4.2	0.08 (0.096)	0.09 (0.028)	**0.60 (0.001)**	**0.23 (0.001)**	**0.13 (0.004)**
Variance of GC content	82.17±23.3	**0.18 (0.003)**	0.09 (0.055)	**0.30 (0.001)**	0.10 (0.026)	0.08 (0.022)
Dinucleotides	79.1%*	0.14 (0.018)	0.08 (0.048)	**0.68 (0.001)**	**0.23 (0.001)**	**0.14 (0.002)**
Effective genome size	1.8±0.3	**0.16 (0.003)**	-0.08 (0.966)	**0.44 (0.001)**	0.06 (0.033)	**0.08 (0.009)**
Number of rRNA/genome	2.7±0.6	**0.17 (0.006)**	-0.03 (0.748)	**0.20 (0.009)**	**0.15 (0.002)**	**0.18 (0.001)**
Number of genes/genome	1362±260	**0.34 (0.001)**	**0.18 (0.002)**	-0.11 (0.991)	0.09 (0.036)	**0.42 (0.001)**
Codons	95.2%*	0.07 (0.134)	0.10 (0.032)	**0.69 (0.001)**	**0.21 (0.001)**	**0.13 (0.001)**
Amino acids	95.5%*	0.13 (0.020)	0.09 (0.026)	**0.69 (0.001)**	**0.24 (0.001)**	**0.17 (0.001)**
Acidic to basic amino acids ratio	0.86±0.02	**0.27 (0.001)**	0.05 (0.120)	**0.36 (0.001)**	**0.32 (0.001)**	**0.24 (0.001)**
% of transcriptional factors	$5{\cdot}10^{-3}$ ±$1{\cdot}10^{-3}$	-0.03 (0.805)	0.00 (0.522)	0.21 (0.011)	0.03 (0.123)	0.04 (0.083)
% of classified reads	65±5	**0.18 (0.003)**	-0.02 (0.636)	**0.30 (0.003)**	0.01 (0.273)	0.06 (0.036)
Functional content	21.6%*	**0.49 (0.001)**	–	**1 (0.001)**	**0.25 (0.001)**	**0.19 (0.001)**
Functional diversity	5.5±0.04	**0.21 (0.001)**	0.06 (0.097)	0.09 (0.094)	0.02 (0.709)	0.00 (0.459)
Taxonomic content	62.6%*	**0.49 (0.001)**	**1 (0.001)**	–	**0.31 (0.001)**	**0.52 (0.001)**
Taxonomic diversity	0.86±0.13	**0.17 (0.007)**	**0.44 (0.001)**	-0.24 (1)	**0.17 (0.003)**	**0.30 (0.001)**
All community traits	40.1%*	**0.29 (0.001)**	**0.42 (0.001)**	**0.58 (0.001)**	**0.33 (0.001)**	**0.42 (0.001)**

11.3.2 Differences between the taxonomic and functional contents

Sequence identity of the 16S rRNA gene has been shown to be related to the overall genomic content in individual genomes (Zaneveld et al., 2010). However, we observed substantial differences between taxonomic (based on the 16S rRNA gene) and functional (based on SEED subsystems) contents in our samples (Mantel test: $r_M = 0.36$, p-value < 0.01; Figure 11.1). Taxonomic content primarily separated the non-oceanic samples (i.e. hypersaline, mangrove, freshwater and estuaries) from the marine plankton (Figure 11.1A), while functional content distinguished communities with different metabolic potentials (Figure 11.1B). The single sample from a harbour (GS149) provides a striking example of how taxonomic community composition and functional content can provide different perspectives of the same complex microbial assemblage. While taxonomically the harbour metagenome was closer to other marine samples, in terms of its functional content it was a unique sample separated from the rest (Figure 11.1A, B). This observation may suggest new research on genomic adaptation in polluted environments and on the dynamic processes that shape microbial communities. For microbial ecologists, it is still an unsolved question whether communities adapt more efficiently modifying the genomic repertoire of their members (as has been shown under laboratory conditions; Sniegowski et al., 1997) or changing the taxonomic composition by ecological processes such as immigration and dispersal.

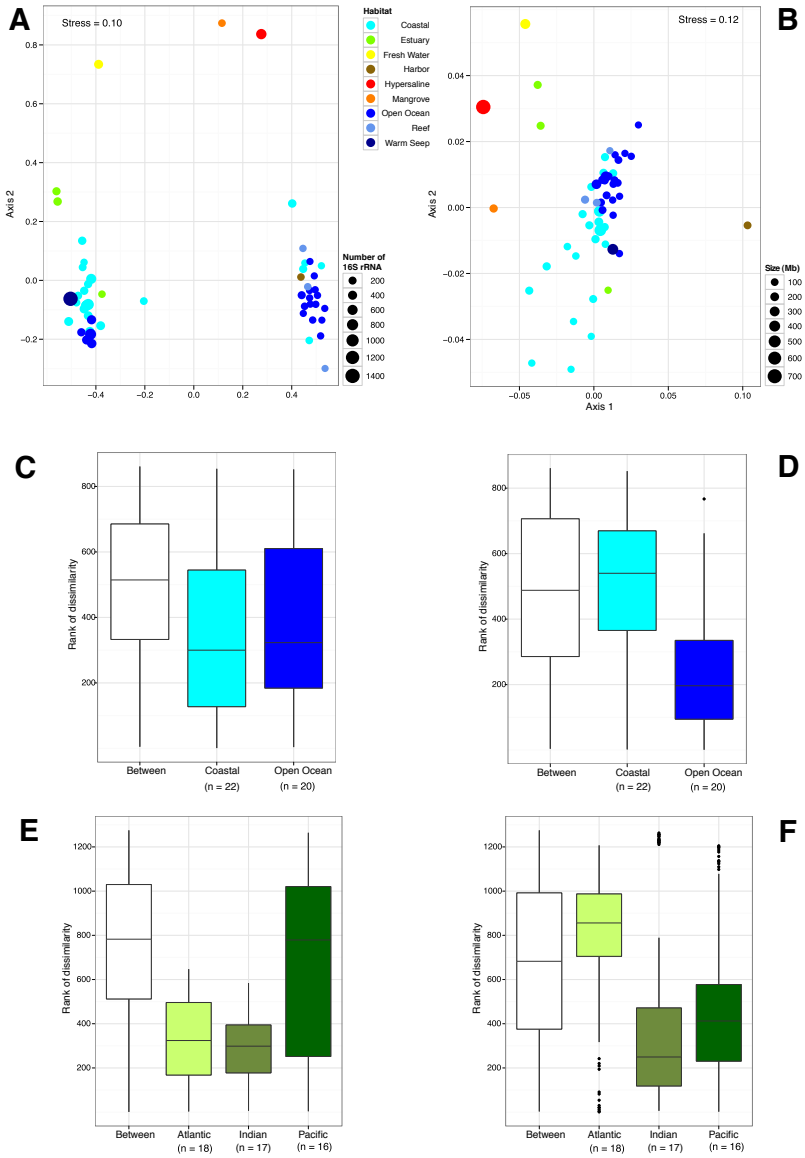


Figure 11.1: Differences between taxonomic and functional community similarity matrices (β-diversity patterns). (*A and B*) Nonmetric multidimensional ordination plots for the taxonomic and functional matrices, respectively. Stress values are indicated. (*C and D*) Rank of dissimilarities between groups and within marine ecosystems (Coastal vs. Open ocean) for the taxonomic and functional matrices, respectively. (*E and F*) Rank of dissimilarities between groups and within oceans (Atlantic vs. Indian vs. Pacific) for the taxonomic and functional matrices, respectively.

Estuaries are productive habitats at the interface of terrestrial and oceanic ecosystems where a mixture of freshwater and marine-specific microorganisms is present (Bouvier & del Giorgio, 2002; Crump et al., 2004). Estuarine samples GS11 (Delaware Bay) and GS12 (Chesapeake Bay) were intermediate between the single freshwater metagenomic sample (GS20) and the marine samples, both taxonomically and functionally (Figure 11.1A, B). Although GS11 and GS12 samples differed in temperature (11 and 3.2 °C, respectively) and chlorophyll concentration (4.8 and 21 $mg{\cdot}m^{-3}$, respectively), they were more similar in composition to each other than to the other estuary sample (GS06 from the Bay of Fundy), which was more similar to other marine metagenomes (Figure 11.1A, B). The temperature and chlorophyll concentration of sample GS06 were very similar to sample GS11. However, other relevant environmental data such as salinity, a key parameter known to greatly affect microbial community composition (Lozupone & Knight, 2007; Auguet et al., 2010), have not been reported for the GOS samples. Thus, we cannot rule out the possible effects of other unmeasured parameters that may explain the observed patterns in community composition. Additionally, GS06 is an estuary sample that also consistently differed in other community traits (Figure 11.2); for example, it had a very low GC content variance (Figure 11.3B) compared with the other estuary samples.

We observed that both functional and taxonomic community patterns significantly varied between coastal and open ocean waters (ANOSIM R = 0.25 and 0.31, respectively). Coastal and open-ocean sites contain water masses with contrasting physicochemical characteristics, and several studies have shown different microbial composition (Acinas et al., 1997; Baltar et al., 2007). Although the ANOSIM R values were similar, functional content clustered all the open-ocean samples together (Figure 11.1C, D). Taxonomic content differentiated better among oceans (ANOSIM R = 0.52) than the functional content (R = 0.19). Taxonomically, the samples from the Pacific Ocean were more heterogeneous (a few samples were more similar to the Atlantic and others to the Indian Ocean), while functionally, the Atlantic Ocean was the most heterogeneous (Figure 11.1E, F; heterogeneous groups of samples show comparable mean rank of dissimilarities to the "Between" category).

11.3.3 Relationships among community traits

We used a PCA to determine the relationship among the community traits and to test how well they could discriminate among samples from different habitats (Figure 11.2). Previous work on bacterial (Lozupone & Knight, 2007) and archaeal (Auguet et al., 2010) community patterns based on the 16S rRNA gene indicated salinity as the major driving force at the global scale. In the

ordination plot of community traits, the hypersaline (GS33) and freshwater (GS20) samples clustered away from the remaining samples (Figure 11.2). The community traits distinguished better among samples from different habitats (PERMANOVA: $r^2 = 0.49$, p-value < 0.001) than taxonomic composition based on the 16S rRNA gene (PERMANOVA: $r^2 = 0.40$, p-value < 0.001).

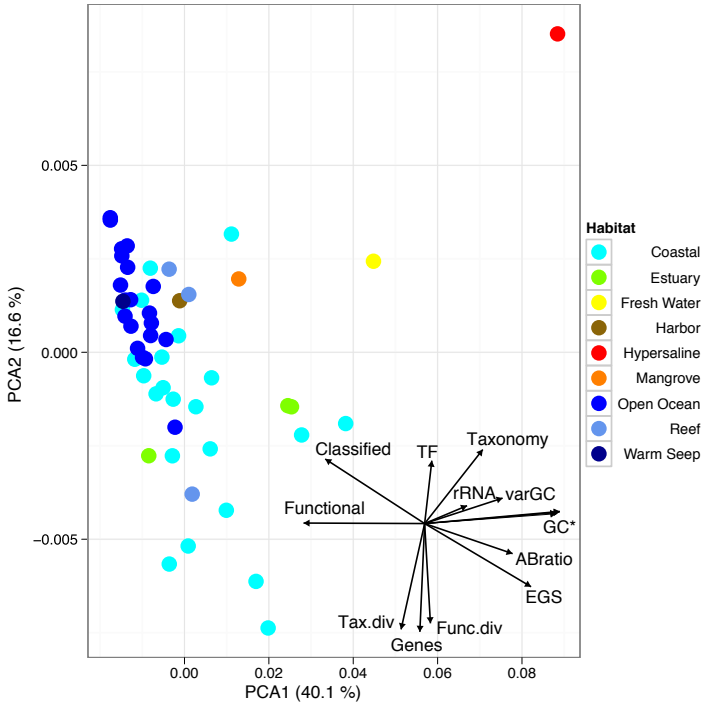


Figure 11.2: Principal component analysis (PCA) ordination plot. In insert, variable loadings centred on (0,0). GC* refers to the highly correlated GC, dinucleotides, codon and aminoacids traits. The variation explained by the two-first components is indicated on the axes.

Assessing bivariate relationships may also help to define ecological strategies across community axes of variation (Wright et al., 2004). A few noteworthy outliers deserve further attention (Figure 11.3). Although a positive relationship (Spearman's $\rho = 0.78$, p-value < 0.01) between the GC content and its variance was detected as a general trend (Figure 11.3B), the hypersaline sample (GS33) clearly deviated, showing a high GC content with low variance. This may reflect a constraining effect of extreme environments at the community level captured in the nucleotide composition, in agreement with the content in the individual genomes reported for hypersaline inhabitants such

as the *Sphingobacteria Salinibacter ruber* (% GC = 66.1) and the *Euryarchaeota Haloquadratum walsbyi* (% GC = 47.9). Further investigations are needed to confirm whether or not this is specifically related to salinity or it can be extrapolated to other extreme environments such as hydrothermal vents or hot springs. Initial analyses using statistical physics methods point to a significant effect of the ecological lifestyle and the composition of functional genes on long-range correlation structure in microbial genomes (Garcia et al., 2008, 2011). Overall, the GC content appeared as a very convenient parameter for initial exploration of metagenomic samples (Foerstner et al., 2005) owing to its straightforward calculation and higher correlation with other properties at higher organizational levels such as dinucleotides, codons, amino acids and functional content (as already known for bacterial genomes) because of the highly dependence on nucleotidic composition (Binnewies et al., 2006, and Figure 11.3A for an example of correlation with the amino acid composition).

A general negative trend (Spearman's ρ = -0.60, p-value < 0.01) between the percentage of TF and EGS (Raes et al., 2007b) was observed (Figure 11.3C). It has been shown that the number of genes in functional categories scales as a power law of the genome size (van Nimwegen, 2003). The freshwater sample (GS20) appeared as an outlier, with a large proportion of TF relative to the remaining metagenomic samples (Figure 11.3C). It has been proposed that TF could be an indicator of environmental variability because transcription factors are more strongly selected in variable than in constant environments (Parter et al., 2007). Lakes are small closed systems, highly diverse and more sensitive to environmental changes than the ocean, and thus promoters of higher microbial diversity (Auguet et al., 2010; Barberán & Casamayor, 2010, 2011; Barberán et al., 2011). More metagenomic samples from different environments and particularly, freshwater samples are certainly needed to confirm this observation.

Another interesting relationship that still needs to be fully explained was the lower percentage of classified reads observed in metagenomic samples with bigger EGS (Figure 11.3D; Spearman's ρ = -0.86, p-value < 0.01). This may result from a truly functional relationship or to sampling bias (the genomes available in public databases may under-represent microorganisms with bigger genomes and larger genomes contain more orphan genes; Skovgaard et al., 2001) or to larger percentage of picoeukaryotes or phages, which are much less characterized.

Finally, rRNA copy number is a trait that had previously attracted considerable attention because it reflects ecological strategies directly related to succession (Klappenbach et al., 2000; Fierer et al., 2007a). At the community level, however, the trait with the highest correlation with rRNA copy number was the ratio of AB amino acids (Spearman's ρ = 0.50, p-value < 0.01). New

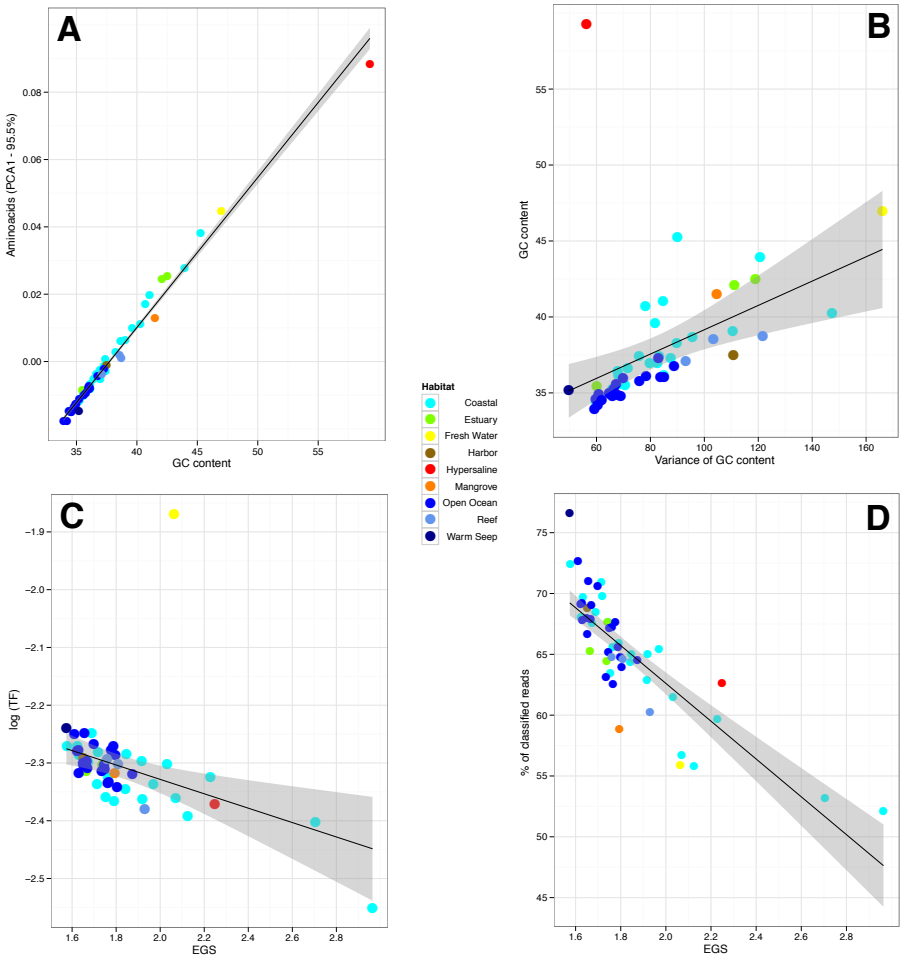


Figure 11.3: Bivariate relationships between some community traits. (*A*) GC content was highly correlated with the aminoacid composition (also with dinucleotide and codon composition). (*B*) Positive relationship between the GC content and its variance. (*C*) Negative relationship between the percentage of transcriptional factors (TF) and the effective genome size (EGS). (*D*) Negative relationship between the percentage of classified reads and the EGS. The general trend is illustrated using a linear regression.

experimental studies should explore how rRNA copy number scales from the population to the community level and how is affected by the environment.

11.3.4 Final conclusions

Overall, the novel approach presented here may help to bridge the gap that exists between the disciplines of general ecology and microbial ecology. The recently developed methodology of metagenomics and trait-based community ecology seems totally compatible and useful for the ecological analysis of complex communities of microorganisms. Although trait-based approaches to microorganisms are largely unexplored (but see recent reviews; Litchman & Klausmeier, 2008; Green et al., 2008), conserved properties at the molecular level (i.e. single-copy genes; Ciccarelli et al., 2006; Wu & Eisen, 2008) and gene length (Xu et al., 2006) serve as anchors to extend the trait approach to the community level in complex environmental genetic data sets.

Acknowledgements

We thank members of the Green laboratory in UO for helpful discussions. AB is supported by the Spanish FPU predoctoral scholarship program and EOC laboratory by grants PIRENA CGL2009-13318 from the Spanish Ministerio de Ciencia e Innovación (MICINN) and the EU-COST Action number ES1103: Microbial Ecology & The Earth System: Collaborating for Insight and Success with the new generation of sequencing tools (CISME).

Appendix [2]

[2]See more Supplementary Information in Barberán et al. (2012b).

Table 11.2: List of the 35 marker genes and their COG (Clusters of Orthologous Groups) annotation.

Orthologous Group	Annotation
COG0012	Predicted GTPase, probable translation factor
COG0016	Phenylalanine-tRNA synthethase alpha subunit
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0085	DNA-directed RNA polymerase, beta subunit
COG0087	Ribosomal protein L3
COG0088	50S ribosomal subunit protein L4
COG0090	50S ribosomal subunit protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6P/L9E
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0124	Histidyl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0185	30S ribosomal subunit protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0256	Ribosomal protein L18
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0525	Valyl-tRNA synthetase
COG0533	Metal-dependent proteases with chaperone activity
COG0541	Signal Recognition Particle (SRP) component with 4.5S RNA (ffs)

General discussion

12
Concluding remarks

Under the constant eager of the Charybdis of fast and voluminous scientific publication, scientists often obviate that as biological entities are shaped by history; every cultural human enterprise, as science itself, also are. Thus, I would like to begin the conclusion of my PhD dissertation with a historical exercise of scientific humility. The best words to summarize most of the results and assertions discussed throughout the different chapters were written more than fifty years ago:

> Experience has shown that a very large number of microbes may be considered to be almost ubiquitous. This does not mean that they are everywhere in considerable numbers, but that a few individuals of the species succeed in maintaining themselves at very divergent spots on earth, either in a dormant state or by temporary and localized outbursts followed by a slow decline of the micropopulation formed. Moreover, many of these germs are of airborne type, and localities temporarily devoid of a certain microbe may soon be repopulated from places where the germ in question has just flourished. Taking into consideration on the one hand the dynamic state of conditions in most soils and waters, implying an almost continuous change in environmental conditions, and on the other hand the marked diversity in the nutritional requirements of various microbial species, it is clear that it is not easy to prophesy which germs will be abundant, which will maintain themselves at a low numerical level, and which will die off in a special locale at a certain moment. (Kluyver & Van Niel, 1956, p. 4-5)

In their book based on a series of lectures at Harvard University in 1954, Albert J. Kluyver (1888-1956) and Cornelius B. van Niel (1897-1985) demonstrated the contribution of microorganisms to genetics and biochemistry, and

discussed the metabolic diversity and the unity of living beings [1]. As the authors responded to the question *What has microbiology offered to general biology?*, the next reasonable question may be *What can microbial ecology offer to general ecology?* Despite some rapprochements during the 20th century (see a recent review in Jessup et al., 2004), ecological studies of microorganisms are historically not part of general ecology, but a subfield of microbiology (O'Malley & Dupré, 2007). With reference to macroecological and biogeographical patterns, it is still uncertain whether life forms show ecological unity, or whether microorganisms and macroorganisms (although similar patterns have been reported) may differ radically in their underlying community processes (Fierer, 2008).

I am convinced that the time for a significant contribution of microbial ecology to general ecology has arrived. Microbial ecology has been as mature as general (i.e. plant and animal) ecology and biogeography was in the 19th century (Fierer, 2008); that is, in a natural history phase. However, it is becoming clear that microbial taxa display non-random environmental and geographical distribution (see recent reviews in Martiny et al., 2006; Dolan, 2006; Ramette & Tiedje, 2007; Fierer, 2008; Soininen, 2012). Most of the patterns that support this evidence have been studied for decades in plant and animal communities but have largely been ignored by microbiologists (notice the time lag of approximately 60 years in Table 2.1 from the **General introduction** section).

Since the first textbook with the term *microbial ecology* in its title (Brock, 1966), the interest in microbial ecology has been undoubtedly triggered by methodological advances, specially in molecular biology. By the beginning of the XXIth century, the relevance (measured as word frequencies in digitized books) of microbial ecology surpassed the one of community ecology (the term *macroecology* has never been very popular since its creation in 1989; Brown & Maurer, 1989), and the study of microbial communities has increased conspicuously since the late 90s (Figure 12.1).

[1] "Anything found to be true of *E. coli* must also be true of elephants" as expressed in 1954 by Jacques Monod, or as coined in 1926 by Kluyver himself "From the elephant to butyric acid bacterium—it is all the same!" (Friedmann, 2004).

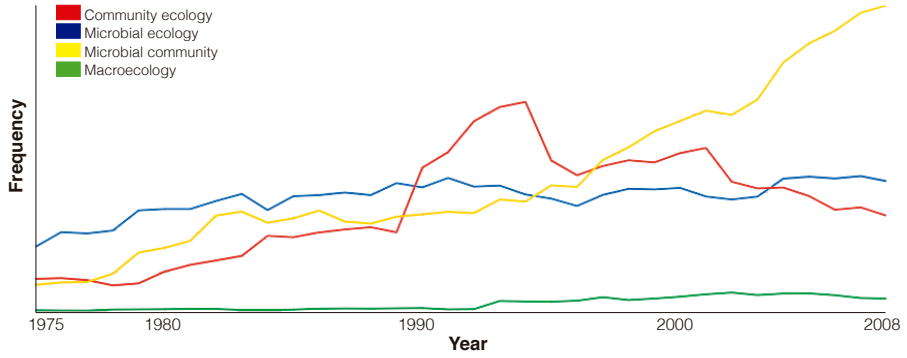


Figure 12.1: Relative frequency in millions of digitized books of key terms for this PhD dissertation. The figure was created using Google Ngrams (http://books.google.com/ngrams; Michel et al., 2011).

Over the past decade microbial ecologists have generated abundant molecular data from ribosomal surveys, and now we are able to combine bioinformatical and statistical tools with critical testing of ecological theory in order to integrate microorganisms into the broader field of ecology (**Chapters 4, 5, 6, 7 and 10**). In parallel from ribosomal surveys, the new field of metagenomics is challenging the scientific community with an astonishing amount of complex data that intersects the disciplines of microbiology, genetics, ecology and bioinformatics. Thus, metagenomics is revealing new complexity of microbial assemblages by moving beyond phylogenetic analyses to functional genes. Comparing the taxonomic and functional community composition may help to clarify whether microbial assemblages change by the adaptation of their genomic repertoire or by the phylogenetic replacement of their members (**Chapter 11**).

In the recent context of climate change, it is of fundamental importance to understand how microbial diversity is altered by environmental modifications already affecting plant and animal diversity (Walther et al., 2002). However microorganisms are frequently neglected from conservational studies, as were in the 1950s and still in the 90s:

> Nowadays most scientists are vaguely aware that something would go wrong if somebody were to succeed in exterminating the microbe world. (Kluyver & Van Niel, 1956, p. 1)

> As one moves down the size-spectrum of organisms, from the romantic large mammals and birds, through nondescript small arthropods, on down to protozoan, bacterial and viral species, not only does concern for diversity and conservation fall away, but it even changes sign. (May, 1990, p. 301)

Accordingly, microbial observatories in pristine and remote environments may serve as sentinels of environmental change, and recorders of long distance atmospheric transportation. A first prospective step of potential mechanisms of microbial dispersal would be the study of the atmospheric deposition travelling long distances at high altitudes and reaching high-mountain lakes (**Chapters 8 and 9**).

Some microbial idiosyncrasies may be of paramount importance for understanding community patterns as described in the **Introduction** section. Comparative studies at large scales have been proven helpful for the testing of different community structures caused by dispersal limitation and environmental filtering (**Chapters 4 and 5**). Furthermore, studies of the atmospheric deposition seem a very convenient initial step to disentangle the potential role of long distance dispersal and survival strategies such as dormancy for community assembly (**Chapter 9**).

Although it is still necessary to accumulate information (i.e. the abovementioned natural history phase) of microbial community composition in the environment (as in **Chapters 8 and 9**), my PhD dissertation has focused on the viability of the transfer of analytical tools from one field to the other. With this objective, my colleagues and I have brought statistical methods and concepts from ecology (e.g. multivariate regression trees in **Chapter 6**, or functional traits in **Chapter 11**), phylogenetic methods from evolutionary biology (e.g. diversification analysis in **Chapter 7**) and network analysis from the science of complex systems (**Chapter 10**) to the field of microbial ecology. Additionally, we have tried to show the potential of the incorporation of phylogenetic information into community ecology (specially in **Chapters 4, 6 and 8**), and we have also proposed a method to describe and analyse metagenomics data, and detect possible anomalies (**Chapter 11**).

In the **General introduction** section, the processes that influence the patterns of community composition, diversity and assembly were compiled as deterministic, stochastic and historical. It has not escaped my notice that throughout my PhD dissertation, deterministic (by means of environmental variables) and historical (through the use of phylogenetic techniques) processes have been given preponderance. Thus, I would desire to conclude this dissertation with a plea for stochasticity.

12.1 ***Coda***: Two neutral models for a unified theory of biodiversity

Understanding the complex and hierarchical structure of biodiversity (the *Baroque of Nature* as expressed by the ecologist Ramon Margalef; Margalef, 1997) is one of the most challenging tasks of modern science (Solé & Bascompte, 2006). In community and ecosystem ecology, the paradigm shift (using Thomas Kuhn's terminology) occured with the transition from a Newtonian metaphysic (deterministic, closed, reversible, universal, and atomistic systems) to an ecological metaphysic (stochastic, open, historical, organic, and hierarchical systems; Ulanowicz, 1999). This dichotomy was already underlying the classical ecological debate between Frederic E. Clements (1874-1945), and Henry A. Gleason (1882-1975). Clements argued that biotic and abiotic factors led to deterministic outcomes of community structure (the *climax state*; Clements, 1916), while Gleason advocated for a certain degree of uncertainty due to the independent behaviour of species (Gleason, 1926).

The classical niche explanation of biological diversity was build upon Gleason's premise of independence, but obviated the role of stochasticity. Niche theory states that every species possesses a unique set of traits that permit its adaptation to a precise abiotic and biotic environment. Conversely, neutral theory focuses on often neglected stochastic processes such as ecological drift and dispersal (which does not mean that it is a form of ecological nihilism). In the words of the proponent of the unified neutral theory of biodiversity and biogeography, Stephen P. Hubbell:

> The dispersal-assembly perspective asserts that ecological communities are open, continuously changing, nonequilibrium assemblages of species whose presence, absence, and relative abundance are governed by random speciation and dispersal, ecological drift, and extinction. (Hubbell, 2001, p. 29)

In ecology, the recognition of chance for community structure dates back to Grinnell (1922). He argued that singletons (that is, accidentals) are the result of fortuitous dispersal. Nevertheless the main source of inspiration for Hubbell was the theory of island biogeography (MacArthur & Wilson, 1967, and see Figure 2.2 for a visualization of the predictions of the theory). If island biogeography was the source of inspiration, the mathematical formulation of the neutral theory in community ecology was derived from population genetics (Leigh Jr, 2007). Kimura (1968) first proposed that the causes of change in allelic frequency are random mutations, migration, and demographic stochasticity. Kimura's neutral theory of molecular evolution has become a null hypothesis by which to test for the presence of natural selection. In this fashion,

Hubbell following an earlier attempt by Caswell (1976) devised a neutral theory for forest dynamics assuming that each tree is equally likely to reproduce or die whatever its species (Hubbell, 1979). His ideas in the form of a developed testable quantitative theory that makes precise predictions across several organizational levels with underlying ecological and evolutionary mechanistic processes arrived in the year 2001 (Hubbell, 2001).

The publication of Hubbell's book raised much controversy among ecologists due to the unrealism and inconsistency with natural history of some of its formulations (Alonso et al., 2006). The most criticized aspect was the assumption of neutrality (i.e. the lack of information about ecological interactions). In Hubbell's model all individuals of different species in a community are strictly equivalent regarding their prospects of reproduction and death. In his neutral framework, the known and evident differences between species are irrelevant for the prediction of large scale patterns, and the crucial point is to determine to what extent functional differences matter. This dispute may be an example of the long-standing philosophical debate between realism (i.e. small scale specific explanations of phenomena) and instrumentalism (i.e. large scale general explanations; Wennekes et al., In press). Along these lines, the simplicity and practicality of ecological models was already acknowledged by the mathematician Alfred J. Lotka (1880-1949):

> There is something unsatisfactory in such abstractions that seem rather far remote from the conditions actually met in nature. But it must be remembered that such abstractions are a necessary, and, as experience has abundantly shown, a very effective aid to our limited mental powers, which are incompetent to deal directly with unexpurgated nature in all its complexity. (Lotka, 1925, p. 301)

Communities may seem neutral because they are complex (i.e. equivalence may occur from non-neutral processes by statistical averaging; Pueyo et al., 2007) with patterns emerging from a statistical process of intricate causalities:

> Communities are not closed systems, so much of what is observed in a local community is determined by processes that transcend the arbitrary boundaries imposed upon it by the observations made on it. This limitation is inescapable. Ecologists simply cannot collect, much less conceptualize, the amount of information that would be necessary to identify precisely why species densities change the ways they do. At the local scale, ecological complexity begets conceptual uncertainty, statistical variability, and theoretical unreliability. (Maurer, 1999, p. 110)

Thus, the neutral theory resembles the kinetic theory of gases: is an ideal theory (Alonso et al., 2006; Etienne & Alonso, 2007). Hubbell brought

parsimony [2] to community ecology: until niche processes are necessary to explain patterns, the simplest neutral models should be considered because *good theory has more predictions per free parameter than does bad theory* (Alonso et al., 2006). Surprisingly, neutral theory predicts observed species abundance distributions, species-area relationships, and β-diversity patterns with distance (Bell, 2001). The falsification of neutral theory, using Karl Popper's central precept, should be against the predictions of models with other mechanistic processes.

Although it may seem that I have been advocating for the neutral theory throughout this conclusion, the depicted model in Hubbell's formulation [3] looks unsatisfactory to a microbial ecologist (though Hubbell's neutral model has already been applied to microbial communities; e.g. Sloan et al., 2006) due to the idiosyncrasies of microbial communities (see section 2.3 for a summary). Here, I propose that a Janus-faced neutral theory composed of two models is required in order to cover the actual biological diversity coarsely divided in macroorganisms and microorganisms (see Figure 12.2 for a conceptual summary). Even if I agree that the same evolutionary and dispersal mechanisms shape species composition, diversity and dynamics of natural communities; there exists a fundamental transition in the scales in which these mechanisms operate (though this may result in similar ecological patterns).

In the modified modelling scheme for microbial communities (Figure 12.2 below), the regional scale is neglected due to high microbial dispersability, and the global scale gains preponderance. If in Hubbell's neutral model formulation for macroorganisms, local and regional scales are coupled through unidirectional migration (i.e. colonization from the metacommunity to the local community); in the microbial model, organisms are allowed to disperse long-distances and thus, be part of the global pool (that is, the rare biosphere or seed bank; Pedrós-Alió, 2006; Lennon & Jones, 2011) as a result of high dispersability and dormancy (Locey, 2010). Hence, individuals may exit local communities by death fueling local extinction, or by long-distance dispersal forming part of the global seed bank. A conceptual procedure to account for higher speciation rates in microorganisms compared to plants and ani-

[2] William of Ockham (c. 1288-c. 1348) is credited as the father of the law of parsimony, known as Occam's razor. The original statement of the English scholastic philosopher was *Frustra fit per plura quod potest fieri per pauciora* (It is futile to do with more what can be done with less).

[3] In Hubbell's model, regional (or metacommunity) dynamics is controlled by the fundamental diversity number (θ), while local dynamics is controlled by the fundamental dispersal number (I). Local dynamics is identical to regional dynamics, except that speciation is replaced by colonization. That is, $\theta \approx 2J_M\nu$ for regional dynamics, and $I \approx 2J_L m$ for local dynamics. Here, J_M represents the number of individuals in the regional pool (i.e. metacommunity size), and J_L the local community size, with $J_L \ll J_M$. The parameters m and ν measure the probability of colonization and speciation, respectively (Hubbell, 2001).

mals (Dykhuizen, 1998) is to allow speciation at the local scale (Figure 12.2). Although both models use the same discrete formulation, the problematic species concept for asexual microorganisms may suggest to treat speciation as a continuous process rather than a discrete one (Rosselló-Mora & Amann, 2001). The study of the derived patterns from such a neutral model focused on microorganisms seem a promising avenue for future research (for another discussion about the plausibility of applying neutral community models to microorganisms see Curtis et al., 2006).

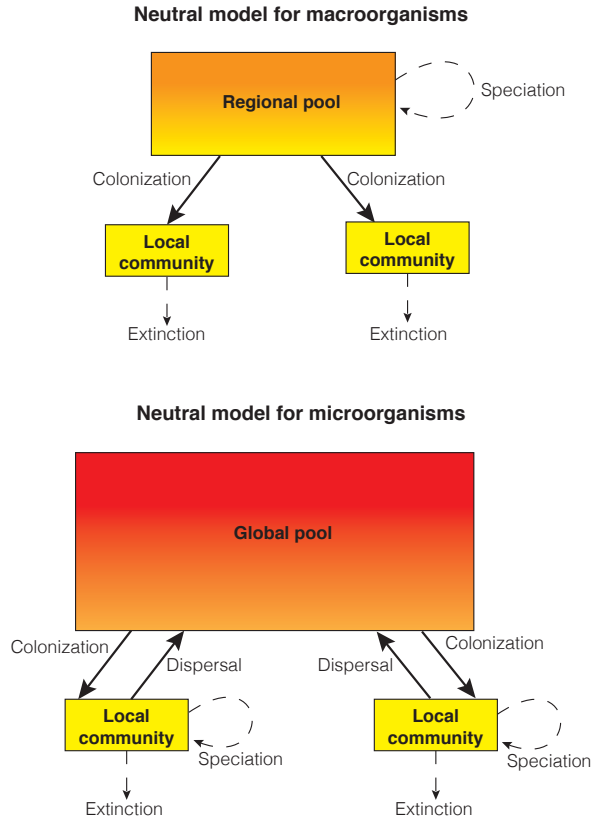


Figure 12.2: A schematic representation of two neutral models: Hubbell's model for macroorganisms (*above*), and a suggested model for microorganisms (*below*). Both models are based on the same mechanistic processes at different scales. See main text for an explanation of the differences between both models. As in Figure 2.1, the demarcation of discrete spatial scales is arbitrary.

In this *coda* I suggested that microbial community ecology should move

beyond Baas Becking's tenet, *everything is everywhere: but the environment selects* (Baas Becking, 1934), towards a stochastic theory built upon evolutionary and ecological mechanisms. Modelling the macrobiological and microbiological neutral models would contribute to the unification of microbial ecology and general ecology:

> It might be noted that even if such a program proves uninteresting the attempt is of value since one measure of our comprehension of ecology is our ability to explain it in a clear unequivocal way to a student with absolutely no prior information about nature or science. An electronic computer is the only such student available. (Slobodkin, 1961, p. 147)

Unveiling ecological patterns and processes from complex and diverse microbial communities is not an easy task. The communities are incredibly diverse with thousands of species, the interactions between these taxa are largely unknown, and the ecological roles of taxa are concealed. Notwithstanding, the enormous diversity and complexity of microbial communities is their key characteristic for the macroecological program (see section 2.1). Again, as in the beginning of this conclusion, looking backwards in time make us realize that we are standing on the shoulders of giants [4]:

> Where microorganisms are concerned, the *islands* can be created artificially [...]. Such synthetic studies can reveal much of general significance but of course nothing about peculiarities in the dispersal and colonizing behavior of the higher plants and animals which comprise most of the several millions of living species on earth. (MacArthur & Wilson, 1967, p. 182)

MacArthur and Wilson acknowledged that microbial communities conform the best living systems to test ecological theories. They also noticed the fundamental divide between macroorganisms and microorganisms. However their statement suffers from an obvious flaw: biological diversity on Earth is mostly microbial. Thus, ecologists cannot simply ignore microbial communities in the development of ecological theory now that we possess the methodological tools to interrogate this long hidden face of diversity.

[4]The metaphor is attributed to the twelfth-century French scholar Bernard of Chartres (in Latin: *Nani gigantum humeris insidentes*) but it was famously reproduced in 1676 by Sir Isaac Newton in a letter to Robert Hooke:

> What Descartes did was a good step. You have added much several ways, and specially in taking the colours of thin plates into philosophical consideration. If I have seen further it is by standing on the shoulders of Giants.

13
Conclusions

The general conclusions of this PhD dissertation are:

- Three main processes explain community patterns: deterministic, stochastic and historic. To account for the historic component of living beings it is essential to incorporate the phylogenetic information in community ecology analysis (**General introduction**).
- Some idiosyncratic features of microorganisms (i.e. high speciation, high dispersability, and dormancy) are of paramount importance in order to understand community patterns (**General introduction**).
- Isolated and environmentally heterogeneous water masses show high levels of microbial diversity (**Chapters 4 and 5**).
- At the global scale and at broad taxonomic levels, environmental filtering is more relevant than geographic processes for explaining the structure of microbial communities (**Chapters 4, 5 and 6**).
- Archaeal communities show definable patterns at the global scale (**Chapter 6**).
- The study of uncultured microbial clades is enhanced by the incorporation of the temporal pattern of diversification (**Chapter 7**).
- Freshwater ecosystems are potential environments for the discovery of new microbial diversity (**Chapters 4, 5, 6 and 7**). In particular, environmental heterogeneity promotes high phylogenetic diversity in Pyrenean lakes (**Chapter 8**).

- The phylogenetic diversity of freshwater bacterial communities from Pyrenean lakes conforms to some biogeographical patterns (**Chapter 8**).
- Dust deposition links global and regional scales of microbial communities in a regular temporal basis (**Chapter 9**).
- Network analysis permits an initial exploration of the co-occurrence patterns among the biotic component (**Chapter 10**) and functional trait analysis allows the summary, description, interpretation and quality assessment of vast metagenomic datasets (**Chapter 11**).
- The taxonomic and functional components show different pictures of microbial communities. Their comparison may help to solve whether communities adapt more efficiently by genetic or by taxonomic variation (**Chapter 11**).
- A neutral community model that neglects the regional scale, considers long-distance dispersal, and locates speciation at the local scale seems more appropriate for microorganisms than previous formulations (**Concluding remarks**).

Bibliography

Bibliography

ACINAS, S., KLEPAC-CERAJ, V., HUNT, D., PHARINO, C., CERAJ, I., DISTEL, D. & POLZ, M., 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, **430**(6999), 551–554.

ACINAS, S., RODRÍGUEZ-VALERA, F. & PEDRÓS-ALIÓ, C., 1997. Spatial and temporal variation in marine bacterioplankton diversity as shown by RFLP fingerprinting of PCR amplified 16S rDNA. *FEMS Microbiology Ecology*, **24**(1), 27–40.

ACKERLY, D., 2009. Phylogenetic Methods in Ecology. *Encyclopedia of Life Sciences (ELS)*.

ACKERT, L., 2007. The Cycle of Life in Ecology: Sergei Vinogradski's Soil Microbiology, 1885–1940. *Journal of the History of Biology*, **40**(1), 109–145.

AGOGUE, H., CASAMAYOR, E., BOURRAIN, M., OBERNOSTERER, I., JOUX, F., HERNDL, G. & LEBARON, P., 2005. A survey on bacteria inhabiting the sea surface microlayer of coastal ecosystems. *FEMS Microbiology Ecology*, **54**(2), 269–280.

ALONSO, D., ETIENNE, R. & MCKANE, A., 2006. The merits of neutral theory. *Trends in Ecology & Evolution*, **21**(8), 451–457.

ALONSO-SÁEZ, L., BALAGUE, V., SA, E., SANCHEZ, O., GONZÁLEZ, J., PINHASSI, J., MASSANA, R., PERNTHALER, J., PEDRÓS-ALIÓ, C. & GASOL, J., 2007. Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiology Ecology*, **60**(1), 98–112.

ANDERSON, M., 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**(1), 32–46.

ANDERSON, M., ELLINGSEN, K. & MCARDLE, B., 2006. Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, **9**(6), 683–693.

ARRHENIUS, O., 1921. Species and area. *Journal of Ecology*, **9**(1), 95–99.

AUGUET, J. C., BARBERÁN, A. & CASAMAYOR, E. O., 2010. Global ecological patterns in uncultured Archaea. *The ISME Journal*, **4**(2), 182–190.

AUGUET, J. C., BORREGO, C. M., BAÑERAS, L. & CASAMAYOR, E. O., 2008. Fingerprinting the genetic diversity of the biotin carboxylase gene (accC) in aquatic ecosystems as a potential marker for studies of carbon dioxide assimilation in the dark. *Environmental Microbiology*, **10**(10), 2527–2536.

AUGUET, J. C. & CASAMAYOR, E. O., 2008. A hotspot for cold crenarchaeota in the neuston of high mountain lakes. *Environmental Microbiology*, **10**(4), 1080–1086.

AUGUET, J. C., TRIADÓ-MARGARIT, X., NOMOKONOVA, N., CAMARERO, L. & CASAMAYOR, E. O., In press. Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophic lake. *The ISME Journal*, doi:10.1038/ismej.2012.33.

BAAS BECKING, L., 1934. *Geobiologie of inleiding tot de milieukunde*. Van Stockum & Zoon, The Hague, the Netherlands.

BALTAR, F., ARÍSTEGUI, J., GASOL, J., HERNÁNDEZ-LEÓN, S. & HERNDL, G., 2007. Strong coast-ocean and surface-depth gradients in prokaryotic assemblage structure and activity in a coastal transition zone region. *Aquatic Microbial Ecology*, **50**(1), 63–74.

BARBERÁN, A., BATES, S. T., CASAMAYOR, E. O. & FIERER, N., 2012a. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, **6**(2), 343–351.

BARBERÁN, A. & CASAMAYOR, E. O., 2010. Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. *Aquatic Microbial Ecology*, **59**(1), 1–10.

BARBERÁN, A. & CASAMAYOR, E. O., 2011. Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microbial Ecology*, **61**(2), 465–472.

BARBERÁN, A., FERNÁNDEZ-GUERRA, A., AUGUET, J. C., GALAND, P. E. & CASAMAYOR, E. O., 2011. Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota. *Molecular Ecology*, **20**(9), 1988–1996.

BARBERÁN, A., FERNÁNDEZ-GUERRA, A., BOHANNAN, B. J. & CASAMAYOR, E. O., 2012b. Exploration of community traits as ecological markers in microbial metagenomes. *Molecular Ecology*, **21**(8), 1909–1917.

BARBOSA, H., MORETTI, M., THULER, D. & AUGUSTO, E., 2002. Nitrogenase activity of *Beijerinckia derxii* is preserved under adverse conditions for its growth. *Brazilian Journal of Microbiology*, **33**(3), 223–229.

BARRACLOUGH, T. & NEE, S., 2001. Phylogenetics and speciation. *Trends in Ecology & Evolution*, **16**(7), 391–399.

BASTIAN, M., HEYMANN, S. & JACOMY, M., 2009. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. San José, California, USA.

BATES, S. T., BERG-LYONS, D., CAPORASO, J. G., WALTERS, W. A., KNIGHT, R. & FIERER, N., 2010. Examining the global distribution of dominant archaeal populations in soil. *The ISME Journal*, **5**(5), 908–917.

BEISNER, B., PERES-NETO, P., LINDSTRÖM, E., BARNETT, A. & LONGHI, M., 2006. The role of environmental and spatial processes in structuring lake communities from bacteria to fish. *Ecology*, **87**(12), 2985–2991.

BELL, G., 2001. Neutral macroecology. *Science*, **293**(5539), 2413–2418.

BENJAMINI, Y. & HOCHBERG, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 289–300.

BENLLOCH, S., LÓPEZ-LÓPEZ, A., CASAMAYOR, E., ØVREÅS, L., GODDARD, V., DAAE, F., SMERDON, G., MASSANA, R., JOINT, I., THINGSTAD, F. ET AL., 2002. Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environmental Microbiology*, **4**(6), 349–360.

BERGMANN, G., BATES, S., EILERS, K., LAUBER, C., CAPORASO, J., WALTERS, W., KNIGHT, R. & FIERER, N., 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology and Biochemistry*, **43**, 1450–1455.

BIDDANDA, B., OGDAHL, M. & COTNER, J., 2001. Dominance of bacterial metabolism in oligotrophic relative to eutrophic waters. *Limnology and Oceanography*, 730–739.

BIERS, E., SUN, S. & HOWARD, E., 2009. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Applied and Environmental Microbiology*, **75**(7), 2221–2229.

BINNEWIES, T., MOTRO, Y., HALLIN, P., LUND, O., DUNN, D., LA, T., HAMPSON, D., BELLGARD, M., WASSENAAR, T. & USSERY, D., 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & Integrative Genomics*, **6**(3), 165–185.

BINTRIM, S., DONOHUE, T., HANDELSMAN, J., ROBERTS, G. & GOODMAN, R., 1997. Molecular phylogeny of archaea from soil. *Proceedings of the National Academy of Sciences USA*, **94**(1), 277–282.

BOUCHER, D., JARDILLIER, L. & DEBROAS, D., 2006. Succession of bacterial community composition over two consecutive years in two aquatic systems: a natural lake and a lake-reservoir. *FEMS Microbiology Ecology*, **55**(1), 79–97.

BOUVIER, T. & DEL GIORGIO, P., 2002. Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. *Limnology and Oceanography*, **47**(2), 453–470.

BOVALLIUS, A., BUCHT, B., ROFFEY, R. & ANÄS, P., 1978. Long-range air transmission of bacteria. *Applied and Environmental Microbiology*, **35**(6), 1231–1232.

BOWERS, R., LAUBER, C., WIEDINMYER, C., HAMADY, M., HALLAR, A., FALL, R., KNIGHT, R. & FIERER, N., 2009. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. *Applied and Environmental Microbiology*, **75**(15), 5121–5130.

BOWERS, R., MCLETCHIE, S., KNIGHT, R. & FIERER, N., 2011a. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *The ISME Journal*, **5**(4), 601–612.

BOWERS, R., SULLIVAN, A., COSTELLO, E., COLLETT JR, J., KNIGHT, R. & FIERER, N., 2011b. Sources of bacteria in outdoor air across cities in the Midwestern United States. *Applied and Environmental Microbiology*, **77**(18), 6350–6356.

BRAY, J. & CURTIS, J., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**(4), 325–349.

BROCHIER-ARMANET, C., BOUSSAU, B., GRIBALDO, S. & FORTERRE, P., 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, **6**(3), 245–252.

BROCK, T. D., 1966. *Principles of Microbial Ecology*. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

BROWN, J. H., 1984. On the relationship between abundance and distribution of species. *The American Naturalist*, **124**(2), 255–279.

BROWN, J. H., 1995. *Macroecology*. University of Chicago Press, Chicago, USA.

BROWN, J. H. & MAURER, B. A., 1989. Macroecology: the division of food and space among species on continents. *Science*, **243**(4895), 1145–1150.

BRYANT, J. A., LAMANNA, C., MORLON, H., KERKHOFF, A. J., ENQUIST, B. J. & GREEN, J. L., 2008. Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences USA*, **105**(Suppl. 1), 11505–11511.

BUENO-HERNÁNDEZ, A. & LLORENTE-BOUSQUETS, J., 2006. The other face of Lyell: historical biogeography in his Principles of geology. *Journal of Biogeography*, **33**(4), 549–559.

BURROWS, S., BUTLER, T., JÖCKEL, P., TOST, H., KERKWEG, A., PÖSCHL, U. & LAWRENCE, M., 2009. Bacteria in the global atmosphere- Part 2: Modelling of emissions and transport between different ecosystems. *Atmospheric Chemistry and Physics Discussions*, **9**(3), 10829–10881.

CAIN, S. A., 1944. *Foundations of Plant Geography*. Hafner Publishing Co., New York, USA.

CAPORASO, J., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F., COSTELLO, E., FIERER, N., PEÑA, A., GOODRICH, J., GORDON, J. ET AL., 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**(5), 335–336.

CASAMAYOR, E. O. & BORREGO, C. M., 2009. *Encyclopedia of Inland Waters*, vol. 3, chap. Archaea in inland waters, 167–181. Elsevier Academic Press.

CASAMAYOR, E. O., CALDERON-PAZ, J. I. & PEDROS-ALIO, C., 2000a. 5S rRNA fingerprints of marine bacteria, halophilic archaea and natural prokaryotic assemblages along a salinity gradient. *FEMS Microbiology Ecology*, **34**(2), 113–119.

CASAMAYOR, E. O., MUYZER, G. & PEDROS-ALIO, C., 2001. Composition and temporal dynamics of planktonic archaeal assemblages from anaerobic sulfurous environments studied by 16S rDNA denaturing gradient gel electrophoresis and sequencing. *Aquatic Microbial Ecology*, **25**(3), 237–246.

CASAMAYOR, E. O., PEDROS-ALIO, C., MUYZER, G. & AMANN, R., 2002. Microheterogeneity in 16S ribosomal DNA-defined bacterial populations from a stratified planktonic environment is related to temporal changes and to ecological adaptations. *Applied and Environmental Microbiology*, **68**(4), 1706–1714.

CASAMAYOR, E. O., SCHAFER, H., BANERAS, L., PEDROS-ALIO, C. & MUYZER, G., 2000b. Identification of and spatio-temporal differences between microbial assemblages from two neighboring sulfurous lakes: Comparison by microscopy and denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology*, **66**(2), 499–508.

CASTRESANA, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**(4), 540.

CASWELL, H., 1976. Community structure: a neutral model analysis. *Ecological Monographs*, **46**, 327–354.

CATALAN, J., BARBIERI, M., BARTUMEUS, F., BITUSIK, P., BOTEV, I., BRANCELJ, A., COGALNICEANU, D., MANCA, M., MARCHETTO, A., OGNJANOVA-RUMENOVA, N. ET AL., 2009. Ecological thresholds in European alpine lakes. *Freshwater Biology*, **54**(12), 2494–2517.

CATALAN, J., CAMARERO, L., FELIP, M., PLA, S., VENTURA, M., BUCHACA, T., BARTUMEUS, F., DE MENDOZA, G., MIRÓ, A., CASAMAYOR, E. O. ET AL., 2006. High mountain lakes: extreme habitats and witnesses of environmental changes. *Limnetica*, **25**(1-2), 551–584.

CATALAN, J., CAMARERO, L., GACIA, E., BALLESTEROS, E. & FELIP, M., 1994. Nitrogen in the Pyrenean lakes (Spain). *Hydrobiologia*, **274**(1), 17–27.

CATALAN, J., VENTURA, M., BRANCELJ, A., GRANADOS, I., THIES, H., NICKUS, U., KORHOLA, A., LOTTER, A., BARBIERI, A., STUCHLÍK, E. ET AL., 2002. Seasonal ecosystem variability in remote mountain lakes: implications for detecting climatic signals in sediment records. *Journal of Paleolimnology*, **28**(1), 25–46.

CAVENDER-BARES, J., KOZAK, K. H., FINE, P. V. & KEMBEL, S. W., 2009. The merging of community ecology and phylogenetic biology. *Ecology Letters*, **12**(7), 693–715.

CHABAN, B., NG, S. & JARRELL, K., 2006. Archaeal habitats-from the extreme to the ordinary. *Canadian Journal of Microbiology*, **52**(2), 73–116.

CHAFFRON, S., REHRAUER, H., PERNTHALER, J. & VON MERING, C., 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, **20**(7), 947–959.

CHASE, J., 2003. Community assembly: when should history matter? *Oecologia*, **136**(4), 489–498.

CHASE, J. & LEIBOLD, M., 2003. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, Chicago, USA.

CHESSON, P. & WARNER, R., 1981. Environmental variability promotes coexistence in lottery competitive systems. *The American Naturalist*, **117**, 923–943.

CHESSON, P. L., 2000. Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, **31**, 343–366.

CICCARELLI, F., DOERKS, T., VON MERING, C., CREEVEY, C., SNEL, B. & BORK, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765), 1283–1287.

CLARKE, K., 1993. Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**(1), 117–143.

CLEMENTS, F., 1916. *Plant succession: an analysis of the development of vegetation*. Carnegie Institution of Washington, Washington DC, USA.

COHAN, F. & KOEPPEL, A., 2008. The origins of ecological diversity in prokaryotes. *Current Biology*, **18**(21), R1024–R1034.

COLEMAN, M., SULLIVAN, M., MARTINY, A., STEGLICH, C., BARRY, K., DELONG, E. & CHISHOLM, S., 2006. Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, **311**(5768), 1768–1770.

CONNELL, J. & ORIAS, E., 1964. The ecological regulation of species diversity. *The American Naturalist*, **98**, 399–414.

COOLEN, M. & OVERMANN, J., 2007. 217,000-year-old DNA sequences of green sulfur bacteria in Mediterranean sapropels and their implications for the reconstruction of the paleoenvironment. *Environmental Microbiology*, **9**(1), 238–249.

COSTELLO, E., LAUBER, C., HAMADY, M., FIERER, N., GORDON, J. & KNIGHT, R., 2009. Bacterial community variation in human body habitats across space and time. *Science*, **326**(5960), 1694–1697.

COTTENIE, K., 2005. Integrating environmental and spatial processes in ecological community dynamics. *Ecology Letters*, **8**(11), 1175–1182.

COTTENIE, K. & DE MEESTER, L., 2004. Metacommunity structure: Synergy of biotic interactions as selective agents and dispersal as fuel. *Ecology*, **85**(1), 114–119.

CRUMP, B., HOPKINSON, C., SOGIN, M. & HOBBIE, J., 2004. Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. *Applied and Environmental Microbiology*, **70**(3), 1494–1505.

CSARDI, G. & NEPUSZ, T., 2006. The igraph software package for complex network research. *InterJournal Complex Systems*, **1695**.

CURTIS, T., HEAD, I., LUNN, M., WOODCOCK, S., SCHLOSS, P. & SLOAN, W., 2006. What is the extent of prokaryotic diversity? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**(1475), 2023–2037.

CURTIS, T., SLOAN, W. & SCANNELL, J., 2002. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences USA*, **99**(16), 10494–10499.

DARLINGTON JR, P., 1959. Area, climate, and evolution. *Evolution*, **13**, 488–510.

DARWIN, C., 1846. An account of the Fine Dust which often falls on Vessels in the Atlantic Ocean. *Quarterly Journal of the Geological Society*, **2**(1-2), 26–30.

DARWIN, C., 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, UK.

DE LA TORRE, J. R., WALKER, C. B., INGALLS, A. E., KOENNEKE, M. & STAHL, D. A., 2008. Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environmental Microbiology*, **10**(3), 810–818.

DE MEESTER, L., 2011. *Biogeography of Microscopic Organisms: Is everything small everywhere?*, chap. A metacommunity perspective on the phylo- and biogegraphy of small organisms, 324–334. Cambridge University Press.

DE WEVER, A., VAN DER GUCHT, K., MUYLAERT, K., COUSIN, S. & VYVERMAN, W., 2008. Clone library analysis reveals an unusual composition and strong habitat partitioning of pelagic bacterial communities in Lake Tanganyika. *Aquatic Microbial Ecology*, **50**(2), 113–122.

DE WIT, R. & BOUVIER, T., 2006. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, **8**(4), 755–758.

DE'ATH, G., 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, **83**(4), 1105–1117.

DELONG, E., 1992. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences USA*, **89**(12), 5685–5689.

DELONG, E., 1998. Everything in moderation: Archaea as 'non-extremophiles'. *Current Opinion in Genetics & Development*, **8**(6), 649–654.

DELONG, E., 2005. Microbial community genomics in the ocean. *Nature Reviews Microbiology*, **3**(6), 459–469.

DEMERGASSO, C., CASAMAYOR, E., CHONG, G., GALLEGUILLOS, P., ESCUDERO, L. & PEDROS-ALIO, C., 2004. Distribution of prokaryotic genetic diversity in athalassohaline lakes of the Atacama Desert, Northern Chile. *FEMS Microbiology Ecology*, **48**(1), 57–69.

DEMERGASSO, C., ESCUDERO, L., CASAMAYOR, E. O., CHONG, G., BALAGUE, V. & PEDROS-ALIO, C., 2008. Novelty and spatio-temporal heterogeneity in the bacterial diversity of hypersaline Lake Tebenquiche (Salar de Atacama). *Extremophiles*, **12**(4), 491–504.

DESANTIS, T. Z., HUGENHOLTZ, P., KELLER, K., BRODIE, E. L., LARSEN, N., PICENO, Y. M., PHAN, R. & ANDERSEN, G. L., 2006a. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, **34**, W394–W399.

DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. & ANDERSEN, G. L., 2006b. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**(7), 5069–5072.

DIAMOND, J., 1975. *Ecology and Evolution of Communities*, chap. Assembly of species communities, 342–444. Harvard University Press, Cambridge, USA.

DINSDALE, E., EDWARDS, R., HALL, D., ANGLY, F., BREITBART, M., BRULC, J., FURLAN, M., DESNUES, C., HAYNES, M., LI, L. ET AL., 2008. Functional metagenomic profiling of nine biomes. *Nature*, **452**(7187), 629–632.

DOLAN, J., 2006. Microbial biogeography? *Journal of Biogeography*, **33**(2), 199–200.

DONACHIE, S., HOU, S., LEE, K., RILEY, C., PIKINA, A., BELISLE, C., KEMPE, S., GREGORY, T., BOSSUYT, A., BOEREMA, J., LIU, J., FREITAS, T., MALAHOFF, A. & ALAM, M., 2004. The Hawaiian Archipelago: A microbial diversity hotspot. *Microbial Ecology*, **48**(4), 509–520.

DOOLITTLE, W. & ZHAXYBAYEVA, O., 2009. On the origin of prokaryotic species. *Genome Research*, **19**(5), 744–756.

DRAKE, J., 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences USA*, **88**(16), 7160–7164.

DUFRENE, M. & LEGENDRE, P., 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs*, **67**(3), 345–366.

DUMESTRE, J. F., CASAMAYOR, E. O., MASSANA, R. & PEDRÓS-ALIÓ, C., 2002. Changes in bacterial and archaeal assemblages in an equatorial river induced by the water eutrophication of Petit Saut dam reservoir (French Guiana). *Aquatic Microbial Ecology*, **26**(3), 209–221.

DWORKIN, M., 2012. Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. *FEMS Microbiology Reviews*, **36**, 364–379.

DYKHUIZEN, D., 1998. Santa Rosalia revisited: Why are there so many species of bacteria? *Antonie van Leeuwenhoek*, **73**, 25–33.

EDGAR, R., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), 2460–2461.

EILER, A. & BERTILSSON, S., 2004. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environmental Microbiology*, **6**(12), 1228–1243.

ELTON, C., 1946. Competition and the structure of ecological communities. *The Journal of Animal Ecology*, **15**, 54–68.

ESTRADA, M., HENRIKSEN, P., GASOL, J., CASAMAYOR, E. & PEDRÓS-ALIÓ, C., 2004. Diversity of planktonic photoautotrophic microorganisms along a salinity gradient as depicted by microscopy, flow cytometry, pigment analysis and DNA-based methods. *FEMS Microbiology Ecology*, **49**(2), 281–293.

ETIENNE, R. S. & ALONSO, D., 2007. Neutral community theory: how stochasticity and dispersal-limitation can explain species coexistence. *Journal of Statistical Physics*, **128**(1), 485–510.

FAITH, D., 1992. Conservation evaluation and phylogenetic diversity. *Biological conservation*, **61**(1), 1–10.

FALKOWSKI, P., FENCHEL, T. & DELONG, E., 2008. The microbial engines that drive Earth's biogeochemical cycles. *Science*, **320**(5879), 1034–1039.

FELIP, M., SATTLER, B., PSENNER, R. & CATALAN, J., 1995. Highly active microbial communities in the ice and snow cover of high mountain lakes. *Applied and Environmental Microbiology*, **61**(6), 2394–2401.

FELSENSTEIN, J., 1985. Phylogenies and the comparative method. *The American Naturalist*, **125**(1), 1–15.

FENCHEL, T., 2003. Biogeography for bacteria. *Science*, **301**(5635), 925–926.

FIELD, D., GARRITY, G., GRAY, T., MORRISON, N., SELENGUT, J. ET AL., 2008. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, **26**(5), 541–547.

FIERER, N., 2008. *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back*, chap. Microbial biogeography: patterns in microbial diversity across space and time, 95–115. ASM Press, Washington DC, USA.

FIERER, N., BRADFORD, M. & JACKSON, R., 2007a. Toward an ecological classification of soil bacteria. *Ecology*, **88**(6), 1354–1364.

FIERER, N., BREITBART, M., NULTON, J., SALAMON, P., LOZUPONE, C., JONES, R., ROBESON, M., EDWARDS, R., FELTS, B., RAYHAWK, S. ET AL., 2007b. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*, **73**(21), 7059–7066.

FIERER, N., HAMADY, M., LAUBER, C. & KNIGHT, R., 2008a. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences USA*, **105**(46), 17994–17999.

FIERER, N. & JACKSON, R., 2006. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences USA*, **103**(3), 626–631.

FIERER, N., LIU, Z., RODRÍGUEZ-HERNÁNDEZ, M., KNIGHT, R., HENN, M. & HERNANDEZ, M., 2008b. Short-term temporal variability in airborne bacterial and fungal populations. *Applied and Environmental Microbiology*, **74**(1), 200–207.

FIERER, N., MCCAIN, C., MEIR, P., ZIMMERMANN, M., RAPP, J., SILMAN, M. & KNIGHT, R., 2011. Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology*, **92**(4), 797–804.

FINLAY, B., 2002. Global dispersal of free-living microbial eukaryote species. *Science*, **296**(5570), 1061–1063.

FISCHER, A., 1960. Latitudinal variations in organic diversity. *Evolution*, **14**(1), 64–81.

FISHER, R., CORBET, A. & WILLIAMS, C., 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, **12**, 42–58.

FOERSTNER, K., VON MERING, C., HOOPER, S. & BORK, P., 2005. Environments shape the nucleotide composition of genomes. *EMBO reports*, **6**(12), 1208–1213.

FORBES, S., 1887. The Lake as a Microcosm. *Bulletin of the Peoria Scientific Association, Illinois*, 77–87.

FRANCIS, C., ROBERTS, K., BEMAN, J., SANTORO, A. & OAKLEY, B., 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences USA*, **102**(41), 14683–14688.

FRANCIS, C. A., BEMAN, J. M. & KUYPERS, M. M. M., 2007. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *The ISME journal*, **1**(1), 19–27.

FRANKLIN, M., MCDONALD, I., BOURNE, D., OWENS, N., UPSTILL-GODDARD, R. & MURRELL, J., 2005. Bacterial diversity in the bacterioneuston (sea surface microlayer): the bacterioneuston through the looking glass. *Environmental Microbiology*, **7**(5), 723–736.

FRASER, C., HANAGE, W. & SPRATT, B., 2007. Recombination and the nature of bacterial speciation. *Science*, **315**(5811), 476–480.

FREILICH, S., KREIMER, A., MEILIJSON, I., GOPHNA, U., SHARAN, R. & RUPPIN, E., 2010. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research*, **38**(12), 3857–3868.

FRETWELL, S., 1975. The impact of Robert MacArthur on ecology. *Annual Review of Ecology and Systematics*, **6**, 1–13.

FRIEDMANN, H., 2004. From “Butyribacterium" to “E. coli": an essay on unity in biochemistry. *Perspectives in Biology and Medicine*, **47**(1), 47–66.

FRIGAARD, N., MARTINEZ, A., MINCER, T. & DELONG, E., 2006. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature*, **439**(7078), 847–850.

FUCHS, B. M., SPRING, S., TEELING, H., QUAST, C., WULF, J., SCHATTENHOFER, M., YAN, S., FERRIERA, S., JOHNSON, J., GLOECKNER, F. O. & AMANN, R., 2007. Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. *Proceedings of the National Academy of Sciences USA*, **104**(8), 2891–2896.

FUHRMAN, J. A., 2009. Microbial community structure and its functional implications. *Nature*, **459**(7244), 193–199.

FUHRMAN, J. A., HEWSON, I., SCHWALBACH, M. S., STEELE, J. A., BROWN, M. V. & NAEEM, S., 2006. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences USA*, **103**(35), 13104–13109.

FUHRMAN, J. A., MCCALLUM, K. & DAVIS, A. A., 1992. Novel major archaebacterial group from marine plankton. *Nature*, **356**(6365), 148–149.

FUHRMAN, J. A. & STEELE, J. A., 2008. Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquatic Microbial Ecology*, **53**(1), 69–81.

FUHRMAN, J. A., STEELE, J. A., HEWSON, I., SCHWALBACH, M. S., BROWN, M. V., GREEN, J. L. & BROWN, J. H., 2008. A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences USA*, **105**(22), 7774–7778.

GALAND, P., CASAMAYOR, E. O., KIRCHMAN, D., POTVIN, M. & LOVEJOY, C., 2009. Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. *The ISME Journal*, **3**(7), 860–869.

GALAND, P., GUTIÉRREZ-PROVECHO, C., MASSANA, R., GASOL, J. & CASAMAYOR, E., 2010. Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnology and Oceanography*, **55**(5), 2117–2125.

GALAND, P., LOVEJOY, C. & VINCENT, W., 2006. Remarkably diverse and contrasting archaeal communities in a large arctic river and the coastal Arctic Ocean. *Aquatic Microbial Ecology*, **44**(2), 115–126.

GALAND, P. E., CASAMAYOR, E. O., KIRCHMAN, D. L. & LOVEJOY, C., 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences USA*, **106**(52), 22427–22432.

GALAND, P. E., POTVIN, M., CASAMAYOR, E. O. & LOVEJOY, C., 2010. Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME journal*, **4**(4), 564–576.

GARCIA, J. A. L., BARTUMEUS, F., ROCHE, D., GIRALDO, J., STANLEY, H. & CASAMAYOR, E. O., 2008. Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses. *Genomics*, **91**(6), 538–543.

GARCIA, J. A. L., FERNÁNDEZ-GUERRA, A. & CASAMAYOR, E. O., 2011. A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes. *Molecular Phylogenetics and Evolution*, **61**(3), 650–658.

GASOL, J. M., CASAMAYOR, E. O., JOINT, I., GARDE, K., GUSTAVSON, K., BENLLOCH, S., DIEZ, B., SCHAUER, M., MASSANA, R. & PEDROS-ALIO, C., 2004. Control of heterotrophic prokaryotic abundance and growth rate in hypersaline planktonic environments. *Aquatic Microbial Ecology*, **34**(2), 193–206.

GAUSE, G. F., 1934. *The Struggle for Existence*. Dover, New York, USA.

GLEASON, H., 1922. On the relation between species and area. *Ecology*, **3**(2), 158–162.

GLEASON, H., 1926. The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club*, **53**, 7–26.

GLISSMAN, K., CHIN, K., CASPER, P. & CONRAD, R., 2004. Methanogenic pathway and archaeal community structure in the sediment of eutrophic Lake Dagow: effect of temperature. *Microbial Ecology*, **48**(3), 389–399.

GLÖCKNER, F. O., FUCHS, B. M. & AMANN, R., 1999. Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. *Applied and Environmental Microbiology*, **65**(8), 3721–3726.

GLÖCKNER, F. O., ZAICHIKOV, E., BELKOVA, N., DENISSOVA, L., PERNTHALER, J., PERNTHALER, A. & AMANN, R., 2000. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology*, **66**(11), 5053–5065.

GOULD, S. J., 1989. *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton & Co., New York, USA.

GOULD, S. J., 2002. *The Structure of Evolutionary Theory*. Belknap Press of Harvard University Press, Cambridge, USA.

GREEN, J., BOHANNAN, B. & WHITAKER, R., 2008. Microbial biogeography: from taxonomy to traits. *Science*, **320**(5879), 1039–1043.

GREEN, J., HOLMES, A., WESTOBY, M., OLIVER, I., BRISCOE, D., DANGERFIELD, M., GILLINGS, M. & BEATTIE, A., 2004. Spatial scaling of microbial eukaryote diversity. *Nature*, **432**(7018), 747–750.

GRIFFIN, D., 2004. Terrestrial microorganisms at an altitude of 20,000 m in Earth's atmosphere. *Aerobiologia*, **20**(2), 135–140.

GRINNELL, J., 1922. The role of the 'accidental'. *The Auk*, **39**(3), 373–380.

GRINNELL, J., 1924. Geography and evolution. *Ecology*, **5**(3), 225–229.

GROSSKOPF, R., STUBNER, S. & LIESACK, W., 1998. Novel euryarchaeotal lineages detected on rice roots and in the anoxic bulk soil of flooded rice microcosms. *Applied and Environmental Microbiology*, **64**(12), 4983–4989.

GUERRERO, R., PIQUERAS, M. & BERLANGA, M., 2002. Microbial mats and the search for minimal ecosystems. *International Microbiology*, **5**(4), 177–188.

GUIMERA, R. & AMARAL, L., 2005. Functional cartography of complex metabolic networks. *Nature*, **433**(7028), 895–900.

GUO, Q., BROWN, J. & VALONE, T., 2000. Abundance and distribution of desert annuals: are spatial and temporal patterns related? *Journal of Ecology*, **88**(4), 551–560.

HAHN, M., 2003. Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones. *Applied and Environmental Microbiology*, **69**(9), 5248–5254.

HANDELSMAN, J., 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**(4), 669–685.

HARDIN, G., 1960. The competitive exclusion principle. *Science*, **131**(3409), 1292–1297.

HELMUS, M. R., BLAND, T. J., WILLIAMS, C. K. & IVES, A. R., 2007. Phylogenetic measures of biodiversity. *The American Naturalist*, **169**(3), E68–E83.

HERFORT, L. & JUNG-HYUN, K., 2009. Diversity of Archaea and detection of crenarchaeotal amoA genes in the rivers Rhine and Tet. *Aquatic Microbial Ecology*, **55**(2), 189–201.

HERMAN, J., BHARTIA, P., TORRES, O., HSU, C., SEFTOR, C. & CELARIER, E., 1997. Global distribution of UV-absorbing aerosols from Nimbus 7/TOMS data. *Journal of Geophysical Research*, **102**(D14), 16911–16922.

HERNDL, G., REINTHALER, T., TEIRA, E., VAN AKEN, H., VETH, C., PERNTHALER, A. & PERNTHALER, J., 2005. Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. *Applied and Environmental Microbiology*, **71**(5), 2303–2309.

HERRERA, C. M., 2005. Resolution of respect. Ramon Margalef (1919-2004). *Bulletin of the Ecological Society of America*, **86**(1), 8–11.

HERUT, B., COLLIER, R. & KROM, M., 2002. The role of dust in supplying nitrogen and phosphorus to the Southeast Mediterranean. *Limnology and Oceanography*, **47**(3), 870–878.

HERVÀS, A., CAMARERO, L., RECHE, I. & CASAMAYOR, E. O., 2009. Viability and potential for immigration of airborne bacteria from Africa that reach high mountain lakes in Europe. *Environmental Microbiology*, **11**(6), 1612–1623.

HERVÀS, A. & CASAMAYOR, E. O., 2009. High similarity between bacterioneuston and airborne bacterial community compositions in a high mountain lake area. *FEMS Microbiology Ecology*, **67**(2), 219–228.

HILLEBRAND, H. & BLENCKNER, T., 2002. Regional and local impact on species diversity-from pattern to processes. *Oecologia*, **132**(4), 479–491.

HOLMES, A., COSTELLO, A., LIDSTROM, M. & MURRELL, J., 1995. Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiology Letters*, **132**(3), 203–208.

HOLYOAK, M., LEIBOLD, M. & HOLT, R., eds., 2005. *Metacommunities: spatial dynamics and ecological communities*. University of Chicago Press, Chicago, USA.

HORNER-DEVINE, M., LAGE, M., HUGHES, J. & BOHANNAN, B., 2004. A taxa-area relationship for bacteria. *Nature*, **432**(7018), 750–753.

HORNER-DEVINE, M., LEIBOLD, M., SMITH, V. & BOHANNAN, B., 2003. Bacterial diversity patterns along a gradient of primary productivity. *Ecology Letters*, **6**(7), 613–622.

HORNER-DEVINE, M. C. & BOHANNAN, B. J. M., 2006. Phylogenetic clustering and overdispersion in bacterial communities. *Ecology*, **87**(7), 100–108.

HORNER-DEVINE, M. C., SILVER, J. M., LEIBOLD, M. A., BOHANNAN, B. J. M., COLWELL, R. K., FUHRMAN, J. A., GREEN, J. L., KUSKE, C. R., MARTINY, J. B. H., MUYZER, G., OVREAS, L., REYSENBACH, A. & SMITH, V. H., 2007. A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology*, **88**(6), 1345–1353.

HOWARD, E., SUN, S., BIERS, E. & MORAN, M., 2008. Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environmental Microbiology*, **10**(9), 2397–2410.

HUBBELL, S. P., 1979. Tree dispersion, abundance, and diversity in a tropical dry forest. *Science*, **203**(4387), 1299–1309.

HUBBELL, S. P., 2001. *The Unified Neutral Theory of Biodiversity and biogeography*. Princeton University Press, Princeton, USA.

HUBER, T., FAULKNER, G. & HUGENHOLTZ, P., 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**(14), 2317–2319.

HUGHES, J., HELLMANN, J., RICKETTS, T. & BOHANNAN, B., 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, **67**(10), 4399–4406.

HULME, M., 2001. Climatic perspectives on Sahelian desiccation: 1973-1998. *Global Environmental Change*, **11**(1), 19–29.

HUMAYOUN, S., BANO, N. & HOLLIBAUGH, J., 2003. Depth distribution of microbial diversity in Mono Lake, a meromictic soda lake in California. *Applied and Environmental Microbiology*, **69**(2), 1030–1042.

HUNT, D. E., DAVID, L. A., GEVERS, D., PREHEIM, S. P., ALM, E. J. & POLZ, M. F., 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, **320**(5879), 1081–1085.

HUTCHINSON, G. E., 1959. Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**(870), 145–159.

HUTCHINSON, G. E., 1961. The paradox of the plankton. *The American Naturalist*, **95**(882), 137–145.

ISHIDA, Y., 2007. Patterns, models, and predictions: Robert MacArthur's approach to ecology. *Philosophy of Science*, **74**(5), 642–653.

IVES, A. & HELMUS, M., 2010. Phylogenetic metrics of community similarity. *The American Naturalist*, **176**(5), E128–E142.

JABLONSKI, D. & SEPKOSKI JR, J., 1996. Paleobiology, community ecology, and scales of ecological pattern. *Ecology*, **77**(5), 1367–1378.

JANSSEN, P., 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology*, **72**(3), 1719–1728.

JESSUP, C., KASSEN, R., FORDE, S., KERR, B., BUCKLING, A., RAINEY, P. & BOHANNAN, B., 2004. Big questions, small worlds: microbial model systems in ecology. *Trends in Ecology & Evolution*, **19**(4), 189–197.

JONES, R., ROBESON, M., LAUBER, C., HAMADY, M., KNIGHT, R. & FIERER, N., 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *The ISME Journal*, **3**(4), 442–453.

JONES, S. & LENNON, J., 2010. Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences USA*, **107**(13), 5881–5886.

JONES, S., NEWTON, R. & MCMAHON, K., 2008. Potential for atmospheric deposition of bacteria to influence bacterioplankton communities. *FEMS Microbiology Ecology*, **64**(3), 388–394.

JURGENS, G., GLOCKNER, F., AMANN, R., SAANO, A., MONTONEN, L., LIKOLAMMI, M. & MUNSTER, U., 2000. Identification of novel Archaea in bacterioplankton of a boreal forest lake by phylogenetic analysis and fluorescent in situ hybridization. *FEMS Microbiology Ecology*, **34**(1), 45–56.

KARNER, M., DELONG, E. & KARL, D., 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature*, **409**(6819), 507–510.

KATOH, K. & TOH, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**(4), 286–298.

KELLOGG, C. & GRIFFIN, D., 2006. Aerobiology and the global transport of desert dust. *Trends in Ecology & Evolution*, **21**(11), 638–644.

KELLY, K. & CHISTOSERDOV, A., 2001. Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island). *FEMS Microbiology Ecology*, **35**(1), 85–95.

KEMBEL, S., COWAN, P., HELMUS, M., CORNWELL, W., MORLON, H., ACKERLY, D., BLOMBERG, S. & WEBB, C., 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**(11), 1463–1464.

KENT, A. D., YANNARELL, A. C., RUSAK, J. A., TRIPLETT, E. W. & MCMAHON, K. D., 2007. Synchrony in aquatic microbial community dynamics. *The ISME journal*, **1**(1), 38–47.

KIMURA, M., 1968. Evolutionary rate at the molecular level. *Nature*, **217**(5129), 624–626.

KLAPPENBACH, J., DUNBAR, J. & SCHMIDT, T., 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, **66**(4), 1328–1333.

KLUYVER, A., 1953. Leeuwenhoek lecture: The changing appraisal of the microbe. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **141**(903), 147–161.

KLUYVER, A. & VAN NIEL, C., 1956. *The Microbe's Contribution to Biology*. Harvard University Press, Cambridge, USA.

KOEPPEL, A., PERRY, E. B., SIKORSKI, J., KRIZANC, D., WARNER, A., WARD, D. M., ROONEY, A. P., BRAMBILLA, E., CONNOR, N., RATCLIFF, R. M., NEVO, E. & COHAN, F. M., 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences USA*, **105**(7), 2504–2509.

KOIZUMI, Y., KOJIMA, H., OGURI, K., KITAZATO, H. & FUKUI, M., 2004. Vertical and temporal shifts in microbial communities in the water column and sediment of saline meromictic Lake Kaiike (Japan), as determined by a 16S rDNA-based analysis, and related to physicochemical gradients. *Environmental Microbiology*, **6**(6), 622–637.

KÖNNEKE, M., BERNHARD, A. E., DE LA TORRE, J. R., WALKER, C. B., WATERBURY, J. B. & STAHL, D. A., 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**(7058), 543–546.

KONOPKA, A., BERCOT, T. & NAKATSU, C., 1999. Bacterioplankton community diversity in a series of thermally stratified lakes. *Microbial Ecology*, **38**(2), 126–135.

KONSTANTINIDIS, K. & TIEDJE, J., 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology*, **10**(5), 504–509.

KRAFT, N., VALENCIA, R. & ACKERLY, D., 2008. Functional traits and niche-based tree community assembly in an Amazonian forest. *Science*, **322**(5901), 580–582.

KRAUSE, A., FRANK, K., MASON, D., ULANOWICZ, R. & TAYLOR, W., 2003. Compartments revealed in food-web structure. *Nature*, **426**, 282–285.

KUCZYNSKI, J., LIU, Z., LOZUPONE, C., MCDONALD, D., FIERER, N. & KNIGHT, R., 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature Methods*, **7**(10), 813–819.

KUNIN, V., COPELAND, A., LAPIDUS, A., MAVROMATIS, K. & HUGENHOLTZ, P., 2008. A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, **72**(4), 557–578.

LANGENHEDER, S. & RAGNARSSON, H., 2007. The role of environmental and spatial factors for the composition of aquatic bacterial communities. *Ecology*, **88**(9), 2154–2161.

LAUBER, C., HAMADY, M., KNIGHT, R. & FIERER, N., 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**(15), 5111–5120.

LEGENDRE, P. & GALLAGHER, E., 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**(2), 271–280.

LEGENDRE, P. & LEGENDRE, L., 1998. *Numerical Ecology*. Developments in Environmental Modeling. Elsevier Science, Amsterdam, the Netherlands.

LEHOURS, A., EVANS, P., BARDOT, C., JOBLIN, K. & GERARD, F., 2007. Phylogenetic diversity of archaea and bacteria in the anoxic zone of a meromictic lake (Lake Pavin, France). *Applied and Environmental Microbiology*, **73**(6), 2016–2019.

LEIBOLD, M., HOLYOAK, M., MOUQUET, N., AMARASEKARE, P., CHASE, J., HOOPES, M., HOLT, R., SHURIN, J., LAW, R., TILMAN, D., LOREAU, M. & GONZALEZ, A., 2004. The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**(7), 601–613.

LEIGH JR, E., 1965. On the relation between the productivity, biomass, diversity, and stability of a community. *Proceedings of the National Academy of Sciences USA*, **53**(4), 777–783.

LEIGH JR, E., 2007. Neutral theory: a historical perspective. *Journal of Evolutionary Biology*, **20**(6), 2075–2091.

LEININGER, S., URICH, T., SCHLOTER, M., SCHWARK, L., QI, J., NICOL, G., PROSSER, J., SCHUSTER, S. & SCHLEPER, C., 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**(7104), 806–809.

LEKUNBERRI, I., LEFORT, T., ROMERO, E., VÁZQUEZ-DOMÍNGUEZ, E., ROMERA-CASTILLO, C., MARRASÉ, C., PETERS, F., WEINBAUER, M. & GASOL, J., 2010. Effects of a dust deposition event on coastal marine microbial abundance and activity, bacterial community structure and ecosystem function. *Journal of Plankton Research*, **32**(4), 381–396.

LENNON, J. & JONES, S., 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*, **9**(2), 119–130.

LETUNIC, I. & BORK, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**(1), 127–128.

LEVIN, S., 1992. The problem of pattern and scale in ecology. *Ecology*, **73**(6), 1943–1967.

LINDSTRÖM, E. & LANGENHEDER, S., 2012. Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**(1), 1–9.

LINDSTRÖM, E. & LOGUE, J., 2008. Biogeography of bacterioplankton in inland waters. *Freshwater Reviews*, **1**(1), 99–114.

LITCHMAN, E. & KLAUSMEIER, C., 2008. Trait-based community ecology of phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 615–639.

LIU, Z., LOZUPONE, C., HAMADY, M., BUSHMAN, F. & KNIGHT, R., 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research*, **35**(18), e120.

LLIRÓS, M., CASAMAYOR, E. O. & BORREGO, C., 2008. High archaeal richness in the water column of a freshwater sulfurous karstic lake along an interannual study. *FEMS Microbiology Ecology*, **66**(2), 331–342.

LOCEY, K., 2010. Synthesizing traditional biogeography with microbial ecology: the importance of dormancy. *Journal of Biogeography*, **37**(10), 1835–1841.

LOGARES, R., BRATE, J., BERTILSSON, S., CLASEN, J., SHALCHIAN-TABRIZI, K. & RENGEFORS, K., 2009. Infrequent marine-freshwater transitions in the microbial world. *Trends in Microbiology*, **17**(9), 414–422.

LOSOS, J., 1996. Phylogenetic perspectives on community ecology. *Ecology*, **77**(5), 1344–1354.

LOSOS, J., 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters*, **11**, 995–1007.

LOTKA, A., 1925. *Elements of Physical Biology*. Williams & Wilkins, New York, USA.

LOZUPONE, C. & KNIGHT, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**(12), 8228–8235.

LOZUPONE, C. A. & KNIGHT, R., 2007. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences USA*, **104**(27), 11436–11440.

LUDWIG, W., STRUNK, O., WESTRAM, R., RICHTER, L., MEIER, H. ET AL., 2004. ARB: a software environment for sequence data. *Nucleic Acids Research*, **32**(4), 1363–1371.

MACARTHUR, R. H., 1965. Patterns of species diversity. *Biological Reviews*, **40**(4), 510–533.

MACARTHUR, R. H., 1972. *Geographical Ecology: patterns in the distribution of species*. Harper & Rowe Publishers, New York, USA.

MACARTHUR, R. H. & LEVINS, R., 1967. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, **101**, 377–385.

MACARTHUR, R. H. & WILSON, E. O., 1967. *The Theory of Island Biogeography*. Princeton University Press, Princeton, USA.

MAGURRAN, A. & HENDERSON, P., 2003. Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**(6933), 714–716.

MALLORQUI, N., ARELLANO, J., BORREGO, C. & GARCIA-GIL, L., 2005. Signature pigments of green sulfur bacteria in lower Pleistocene deposits from the Banyoles lacustrine area (Spain). *Journal of Paleolimnology*, **34**(2), 271–280.

MARGALEF, R., 1963. On certain unifying principles in ecology. *The American Naturalist*, **97**(897), 357–374.

MARGALEF, R., 1968. *Perspectives in Ecological Theory*. University of Chicago Press, Chicago, USA.

MARGALEF, R., 1996. Information and uncertainty in living systems, a view from ecology. *Biosystems*, **38**, 141–146.

MARGALEF, R., 1997. *Our Biosphere*, vol. 10 of *Excellence in Ecology Series*. Ecology Institute Oldendorf/Luhe, Germany.

MARTIN, A., COSTELLO, E., MEYER, A., NEMERGUT, D. & SCHMIDT, S., 2004. The rate and pattern of cladogenesis in microbes. *Evolution*, **58**(5), 946–955.

MARTIN-CUADRADO, A., RODRIGUEZ-VALERA, F., MOREIRA, D., ALBA, J., IVARS-MARTÍNEZ, E., HENN, M., TALLA, E. & LÓPEZ-GARCÍA, P., 2008. Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *The ISME Journal*, **2**(8), 865–886.

MARTINY, J., BOHANNAN, B., BROWN, J., COLWELL, R., FUHRMAN, J., GREEN, J., HORNER-DEVINE, M., KANE, M., KRUMINS, J., KUSKE, C. ET AL., 2006. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**(2), 102–112.

MAURER, B. A., 1999. *Untangling Ecological Complexity: the macroscopic perspective*. University of Chicago Press, Chicago, USA.

MAY, R., 1974. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science*, **186**(4164), 645–647.

MAY, R. M., 1990. How many species? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **330**(1257), 293–304.

MAYNARD SMITH, J., SMITH, N., O'ROURKE, M. & SPRATT, B., 1993. How clonal are bacteria? *Proceedings of the National Academy of Sciences USA*, **90**(10), 4384–4388.

MAYR, E., 1961. Cause and effect in biology. *Science*, **134**(3489), 1501–1506.

MCARDLE, B. & ANDERSON, M., 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, **82**(1), 290–297.

MCCAIG, A., GLOVER, L. & PROSSER, J., 2001. Numerical analysis of grassland bacterial community structure under different land management regimens by using 16S ribosomal DNA sequence data and denaturing gradient gel electrophoresis banding patterns. *Applied and Environmental Microbiology*, **67**(10), 4554–4559.

MCGILL, B., ENQUIST, B., WEIHER, E. & WESTOBY, M., 2006. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, **21**(4), 178–185.

MCPEEK, M., 2008. The ecological dynamics of clade diversification and community assembly. *The American Naturalist*, **172**(6), E270–284.

MESBAH, N., ABOU-EL-ELA, S. & WIEGEL, J., 2007. Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun, Egypt. *Microbial Ecology*, **54**(4), 598–617.

MEYER, F., PAARMANN, D., D'SOUZA, M., OLSON, R., GLASS, E., KUBAL, M., PACZIAN, T., RODRIGUEZ, A., STEVENS, R., WILKE, A. ET AL., 2008. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**(1), 386.

MICHEL, J., SHEN, Y., AIDEN, A., VERES, A., GRAY, M., PICKETT, J., HOIBERG, D., CLANCY, D., NORVIG, P., ORWANT, J. ET AL., 2011. Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014), 176–182.

MLADENOV, N., SOMMARUGA, R., MORALES-BAQUERO, R., LAURION, I., CAMARERO, L., DIÉGUEZ, M. C., CAMACHO, A., DELGADO, A., TORRES, O., CHEN, Z., FELIP, M. & RECHE, I., 2011. Dust inputs and bacteria influence dissolved organic matter in clear alpine lakes. *Nature Communications*, **2**, 405.

MONOD, J., 1971. *Chance and Necessity: an essay on the natural philosophy of modern biology*. Alfred A. Knopf, New York, USA.

MOODY, J., 2001. Race, school integration, and friendship segregation in America. *American Journal of Sociology*, **107**(3), 679–716.

MOOERS, A. & HEARD, S., 1997. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, **72**(1), 31–54.

MORALES-BAQUERO, R., PULIDO-VILLENA, E. & RECHE, I., 2006. Atmospheric inputs of phosphorus and nitrogen to the southwest Mediterranean region: Biogeochemical responses of high mountain lakes. *Limnology and Oceanography*, **51**(2), 830–837.

MORLON, H., SCHWILK, D., BRYANT, J., MARQUET, P., REBELO, A., TAUSS, C., BOHANNAN, B. & GREEN, J., 2011. Spatial patterns of phylogenetic diversity. *Ecology Letters*, **14**(2), 141–149.

MORRIS, R., RAPPÉ, M., CONNON, S., VERGIN, K., SIEBOLD, W., CARLSON, C., GIOVANNONI, S. ET AL., 2002. SAR 11 clade dominates ocean surface bacterioplankton communities. *Nature*, **420**(6917), 806–810.

MOULIN, C. & CHIAPELLO, I., 2006. Impact of human-induced desertification on the intensification of Sahel dust emission and export over the last decades. *Geophysical Research Letters*, **33**(18), L18808.

MUÑOZ, J., FELICÍSIMO, Á., CABEZAS, F., BURGAZ, A. & MARTÍNEZ, I., 2004. Wind as a long-distance dispersal vehicle in the Southern Hemisphere. *Science*, **304**(5674), 1144–1147.

NELSON, C., 2009. Phenology of high-elevation pelagic bacteria: the roles of meteorologic variability, catchment inputs and thermal stratification in structuring communities. *The ISME Journal*, **3**(1), 13–30.

NEWMAN, M. E. J., 2003. The structure and function of complex networks. *SIAM Review*, **45**, 167–256.

NEWMAN, M. E. J., 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences USA*, **103**(23), 8577.

NEWTON, R. J., JONES, S. E., HELMUS, M. R. & MCMAHON, K. D., 2007. Phylogenetic ecology of the freshwater Actinobacteria acI lineage. *Applied and Environmental Microbiology*, **73**(22), 7169–7176.

NEWTON, R. J., KENT, A. D., TRIPLETT, E. W. & MCMAHON, K. D., 2006. Microbial community dynamics in a humic lake: differential persistence of common freshwater phylotypes. *Environmental Microbiology*, **8**(6), 956–970.

NG, I. S. Y., CARR, C. M. & COTTENIE, K., 2009. Hierarchical zooplankton metacommunities: distinguishing between high and limiting dispersal mechanisms. *Hydrobiologia*, **619**, 133–143.

OCHSENREITER, T., SELEZI, D., QUAISER, A., BONCH-OSMOLOVSKAYA, L. & SCHLEPER, C., 2003. Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environmental Microbiology*, **5**(9), 787–797.

ODUM, E. P., 1969. The strategy of ecosystem development. *Science*, **169**, 262–270.

OKSANEN, J., KINDT, R., LEGENDRE, P., OHARA, B., SIMPSON, G., SOLYMOS, P., STEVENS, M. & WAGNER, H., 2009. *vegan: Community Ecology Package. R package version 1.15-4*.

O'MALLEY, M., 2008. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Biological and Biomedical Sciences*, **39**(3), 314–325.

O'MALLEY, M. & DUPRÉ, J., 2007. Size doesn't matter: towards a more inclusive philosophy of biology. *Biology and Philosophy*, **22**(2), 155–191.

ORIANS, G., 1969. The number of bird species in some tropical forests. *Ecology*, **50**, 783–801.

O'SULLIVAN, L., FULLER, K., THOMAS, E., TURLEY, C., FRY, J. & WEIGHTMAN, A., 2004. Distribution and culturability of the uncultivated AGG58 cluster of the Bacteroidetes phylum in aquatic environments. *FEMS Microbiology Ecology*, **47**(3), 359–370.

OVERBEEK, R., BEGLEY, T., BUTLER, R., CHOUDHURI, J., CHUANG, H., COHOON, M., DE CRÉCY-LAGARD, V., DIAZ, N., DISZ, T., EDWARDS, R. ET AL., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, **33**(17), 5691–5702.

OVREAS, L., 2000. Population and community level approaches for analysing microbial diversity in natural environments. *Ecology Letters*, **3**(3), 236–251.

PACE, N. R., 1997. A molecular view of microbial diversity and the biosphere. *Science*, **276**(5313), 734–740.

PADIAN, K. & CHIAPPE, L., 1998. The origin and early evolution of birds. *Biological Reviews*, **73**(1), 1–42.

PANDIT, S., KOLASA, J. & COTTENIE, K., 2009. Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology*, **90**(8), 2253–2262.

PAPKE, R. & WARD, D., 2004. The importance of physical isolation to microbial diversification. *FEMS Microbiology Ecology*, **48**(3), 293–303.

PARADIS, E., CLAUDE, J. & STRIMMER, K., 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**(2), 289–290.

PARTER, M., KASHTAN, N. & ALON, U., 2007. Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, **7**, 169.

PASTOR-SATORRAS, R. & VESPIGNANI, A., 2001. Epidemic spreading in scale-free networks. *Physical Review Letters*, **86**(14), 3200–3203.

PEARCE, D., VAN DER GAST, C., WOODWARD, K. & NEWSHAM, K., 2005. Significant changes in the bacterioplankton community structure of a maritime Antarctic freshwater lake following nutrient enrichment. *Microbiology*, **151**(10), 3237–3248.

PEDRÓS-ALIÓ, C., 2006. Marine microbial diversity: can it be determined? *Trends in Microbiology*, **14**(6), 257–263.

PERNTHALER, J., GLÖCKNER, F., UNTERHOLZNER, S., ALFREIDER, A., PSENNER, R. & AMANN, R., 1998. Seasonal community and population dynamics of pelagic bacteria and archaea in a high mountain lake. *Applied and Environmental Microbiology*, **64**(11), 4299–4306.

PHILIPPOT, L., ANDERSSON, S., BATTIN, T., PROSSER, J., SCHIMEL, J., WHITMAN, W. & HALLIN, S., 2010. The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology*, **8**(7), 523–529.

POMMIER, T., CANBACK, B., RIEMANN, L., BOSTROM, K. H., SIMU, K., LUNDBERG, P., TUNLID, A. & HAGSTROM, A., 2007. Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, **16**(4), 867–880.

POULIOT, J., GALAND, P., LOVEJOY, C. & VINCENT, W., 2009. Vertical structure of archaeal communities and the distribution of ammonia monooxygenase A gene variants in two meromictic High Arctic lakes. *Environmental Microbiology*, **11**(3), 687–699.

PRESTON, F. W., 1948. The commonnes, and rarity, of species. *Ecology*, **29**, 254–283.

PROSPERO, J., BLADES, E., MATHISON, G. & NAIDU, R., 2005. Interhemispheric transport of viable fungi and bacteria from Africa to the Caribbean with soil dust. *Aerobiologia*, **21**(1), 1–19.

PROSPERO, J. & LAMB, P., 2003. African droughts and dust transport to the Caribbean: Climate change implications. *Science*, **302**(5647), 1024–1027.

PROSSER, J., BOHANNAN, B., CURTIS, T., ELLIS, R., FIRESTONE, M., FRECKLETON, R., GREEN, J., GREEN, L., KILLHAM, K., LENNON, J. ET AL., 2007. The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, **5**, 384–392.

PROULX, S., PROMISLOW, D. & PHILLIPS, P., 2005. Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, **20**(6), 345–353.

PSENNER, R., 1999. Living in a dusty world: airborne dust as a key factor for alpine lakes. *Water, Air, & Soil Pollution*, **112**(3), 217–227.

PUEYO, S., HE, F. & ZILLIO, T., 2007. The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecology Letters*, **10**(11), 1017–1028.

PURVIS, A., AGAPOW, P., GITTLEMAN, J. & MACE, G., 2000. Nonrandom extinction and the loss of evolutionary history. *Science*, **288**(5464), 328–330.

PYBUS, O. & HARVEY, P., 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, **267**(1459), 2267–2272.

QUISPEL, A., 1998. Lourens G. M. Baas Becking (1895–1963). Inspirator for many (micro)biologists. *International Microbiology*, **1**, 69–72.

RAES, J., FOERSTNER, K. & BORK, P., 2007a. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology*, **10**(5), 490–498.

RAES, J., KORBEL, J., LERCHER, M., VON MERING, C. & BORK, P., 2007b. Prediction of effective genome size in metagenomic samples. *Genome Biology*, **8**, R10.

RAES, J., LETUNIC, I., YAMADA, T., JENSEN, L. & BORK, P., 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*, **7**, 473.

RAMETTE, A. & TIEDJE, J. M., 2007. Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology*, **53**(2), 197–207.

RECHE, I., ORTEGA-RETUERTA, E., ROMERA, O., PULIDO-VILLENA, E., MORALES-BAQUERO, R. & CASAMAYOR, E., 2009. Effect of Saharan dust inputs on bacterial activity and community composition in Mediterranean lakes and reservoirs. *Limnology and Oceanography*, **54**(3), 869–879.

RECHE, I., PULIDO-VILLENA, E., MORALES-BAQUERO, R. & CASAMAYOR, E. O., 2005. Does ecosystem size determine aquatic bacterial richness? *Ecology*, **86**(7), 1715–1722.

RECHE, I., PULIDO-VILLENA, E., MORALES-BAQUERO, R. & CASAMAYOR, E. O., 2007. Does ecosystem size determine aquatic bacteria richness? Reply. *Ecology*, **88**(1), 253–255.

RHODES, M., FITZ-GIBBON, S., OREN, A. & HOUSE, C., 2010. Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environmental Microbiology*, **12**(9), 2613–2623.

RICE, P., LONGDEN, I., BLEASBY, A. ET AL., 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**(6), 276–277.

RICKLEFS, R. E., 1987. Community diversity: relative roles of local and regional processes. *Science*, **235**, 167–171.

RICKLEFS, R. E., 2008. Disintegration of the Ecological Community. *The American Naturalist*, **172**(6), 741–750.

ROBERTSON, C., HARRIS, J., SPEAR, J. & PACE, N., 2005. Phylogenetic diversity and ecology of environmental Archaea. *Current Opinion in Microbiology*, **8**(6), 638–642.

RODRIGUES, A. & GASTON, K., 2002. Maximising phylogenetic diversity in the selection of networks of conservation areas. *Biological Conservation*, **105**(1), 103–111.

ROESCH, L., FULTHORPE, R., RIVA, A., CASELLA, G., HADWIN, A., KENT, A., DAROUB, S., CAMARGO, F., FARMERIE, W. & TRIPLETT, E., 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**(4), 283–290.

ROSSELLÓ-MORA, R. & AMANN, R., 2001. The species concept for prokaryotes. *FEMS Microbiology Ecology*, **25**(1), 39–67.

RUAN, Q., DUTTA, D., SCHWALBACH, M., STEELE, J., FUHRMAN, J. & SUN, F., 2006. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, **22**(20), 2532–2538.

RUSCH, D., HALPERN, A., SUTTON, G., HEIDELBERG, K., WILLIAMSON, S., YOOSEPH, S., WU, D., EISEN, J., HOFFMAN, J., REMINGTON, K. ET AL., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **5**(3), e77.

SALCHER, M., PERNTHALER, J., ZEDER, M., PSENNER, R. & POSCH, T., 2008. Spatio-temporal niche separation of planktonic Betaproteobacteria in an oligo-mesotrophic lake. *Environmental Microbiology*, **10**(8), 2074–2086.

SALTHE, S., 1985. *Evolving hierarchical systems*. Columbia University Press, New York, USA.

SANDERSON, M., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, **14**(12), 1218–1231.

SCHLEPER, C., 2007. *Archaea: Evolution, Physiology and Molecular Biology*, chap. Diversity of uncultivated Archaea: perspectives from microbial ecology and metagenomics, 39.53. Blackwell Publishing, Oxford, UK.

SCHLEPER, C., JURGENS, G. & JONUSCHEIT, M., 2005. Genomic studies of uncultivated archaea. *Nature Reviews Microbiology*, **3**(6), 479–488.

SCHLOSS, P. & HANDELSMAN, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, **71**(3), 1501–1506.

SCHOPF, J. & PACKER, B., 1987. Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science*, **237**(4810), 70–73.

SCHRÖDINGER, E., 1944. *What is life?*. Cambridge University Press, Cambridge, UK.

SEGERER, A., BURGGRAF, S., FIALA, G., HUBER, G., HUBER, R., PLEY, U. & STETTER, K., 1993. Life in hot-springs and hydrothermal vents. *Origins of Life and Evolution of the Biosphere*, **23**(1), 77–90.

SEKIGUCHI, H., WATANABE, M., NAKAHARA, T., XU, B. & UCHIYAMA, H., 2002. Succession of bacterial community structure along the Changjiang River determined by denaturing gradient gel electrophoresis and clone library analysis. *Applied and Environmental Microbiology*, **68**(10), 5142–5150.

SESHADRI, R., KRAVITZ, S., SMARR, L., GILNA, P. & FRAZIER, M., 2007. CAMERA: a community resource for metagenomics. *PLoS Biology*, **5**(3), e75.

SHADE, A., JONES, S. E. & MCMAHON, K. D., 2008. The influence of habitat heterogeneity on freshwater bacterial community composition and dynamics. *Environmental Microbiology*, **10**(4), 1057–1067.

SHAW, A. K., HALPERN, A. L., BEESON, K., TRAN, B., VENTER, J. C. & MARTINY, J. B. H., 2008. It's all relative: ranking the diversity of aquatic bacterial communities. *Environmental Microbiology*, **10**(9), 2200–2210.

SIBLEY, C. & AHLQUIST, J., 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *Journal of Molecular Evolution*, **26**(1), 99–121.

SIMBERLOFF, D., 1980. A succession of paradigms in ecology: essentialism to materialism and probabilism. *Synthese*, **43**(1), 3–39.

SKOVGAARD, M., JENSEN, L., BRUNAK, S., USSERY, D. & KROGH, A., 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, **17**(8), 425–428.

SLOAN, W., LUNN, M., WOODCOCK, S., HEAD, I., NEE, S. & CURTIS, T., 2006. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, **8**(4), 732–740.

SLOBODKIN, L., 1961. Preliminary ideas for a predictive theory of ecology. *The American Naturalist*, **95**(882), 147–153.

SNIEGOWSKI, P., GERRISH, P., LENSKI, R. ET AL., 1997. Evolution of high mutation rates in experimental populations of E. coli. *Nature*, **387**, 703–705.

SOGIN, M., MORRISON, H., HUBER, J., WELCH, D., HUSE, S., NEAL, P., ARRIETA, J. & HERNDL, G., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences USA*, **103**(32), 12115–12120.

SOININEN, J., 2012. Macroecology of unicellular organisms – patterns and processes. *Environmental Microbiology Reports*, **4**(1), 10–22.

SOININEN, J., KORHONEN, J., KARHU, J. & VETTERLI, A., 2011. Disentangling the spatial patterns in community composition of prokaryotic and eukaryotic lake plankton. *Limnology and Oceanography*, **56**(2), 508–520.

SOLÉ, R. V. & BASCOMPTE, J., 2006. *Self-Organization in Complex Ecosystems*. Princeton University Press, Princeton, USA.

SOMMARUGA, R., 2001. The role of solar UV radiation in the ecology of alpine lakes. *Journal of Photochemistry and Photobiology B: Biology*, **62**(1-2), 35–42.

SOMMARUGA, R. & CASAMAYOR, E. O., 2009. Bacterial 'cosmopolitanism' and importance of local environmental factors for community composition in remote high-altitude lakes. *Freshwater Biology*, **54**(5), 994–1005.

SOUZA, V., ESPINOSA-ASUAR, L., ESCALANTE, A., EGUIARTE, L., FARMER, J., FORNEY, L., LLORET, L., RODRIGUEZ-MARTINEZ, J., SOBERON, X., DIRZO, R. & ELSER, J., 2006. An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proceedings of the National Academy of Sciences USA*, **103**(17), 6565–6570.

SPAIN, A., KRUMHOLZ, L. & ELSHAHED, M., 2009. Abundance, composition, diversity and novelty of soil Proteobacteria. *The ISME Journal*, **3**(8), 992–1000.

STACKEBRANDT, E. & GOEBEL, B., 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology*, **44**(4), 846–849.

STAMATAKIS, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–2690.

STEINHAUSER, D., KRALL, L., MÜSSIG, C., DÜSSIS, D. & USADEL, B., 2008. *Analysis of Biological Networks*, chap. Correlation Networks. Wiley Online Library.

STONE, L. & ROBERTS, A., 1990. The checkerboard score and species distributions. *Oecologia*, **85**(1), 74–79.

TEELING, H., MEYERDIERKS, A., BAUER, M., AMANN, R. & GLÖCKNER, F., 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, **6**(9), 938–947.

TELFORD, R., VANDVIK, V. & BIRKS, H., 2006. Dispersal limitations matter for microbial morphospecies. *Science*, **312**(5776), 1015.

THOMPSON, D. W., 1942. *On Growth and Form*. Cambridge University Press, Cambridge, UK, 2nd ed.

TILMAN, D., 1982. *Resource competition and community structure*. Princeton University Press, Princeton, USA.

TONG, Y. & LIGHTHART, B., 1998. Effect of simulated solar radiation on mixed outdoor atmospheric bacterial populations. *FEMS Microbiology Ecology*, **26**(4), 311–316.

TORSVIK, V., GOKSOYR, J. & DAAE, F., 1990. High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology*, **56**(3), 782–787.

TORSVIK, V. & OVREAS, L., 2002. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology*, **5**(3), 240–245.

TRIADÓ-MARGARIT, X. & CASAMAYOR, E., In press. Genetic diversity of planktonic eukaryotes in high mountain lakes (Central Pyrenees, Spain). *Environmental Microbiology*, doi: 10.1111/j.1462–2920.2012.02797.x.

ULANOWICZ, R., 1999. Life after Newton: an ecological metaphysic. *BioSystems*, **50**(2), 127–142.

URBACH, E., VERGIN, K., YOUNG, L., MORSE, A., LARSON, G. & GIOVANNONI, S., 2001. Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnology and Oceanography*, **46**(3), 557–572.

VAMOSI, S. M., HEARD, S. B., VAMOSI, J. C. & WEBB, C. O., 2009. Emerging patterns in the comparative analysis of phylogenetic community structure. *Molecular Ecology*, **18**(4), 572–592.

VAN DER GAST, C., WALKER, A., STRESSMANN, F., ROGERS, G., SCOTT, P., DANIELS, T., CARROLL, M., PARKHILL, J. & BRUCE, K., 2011. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *The ISME Journal*, **5**, 780–791.

VAN DER GUCHT, K., COTTENIE, K., MUYLAERT, K., VLOEMANS, N., COUSIN, S., DECLERCK, S., JEPPESEN, E., CONDE-PORCUNA, J., SCHWENK, K., ZWART, G., DEGANS, H., VYVERMAN, W. & DE MEESTER, L., 2007. The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proceedings of the National Academy of Sciences USA*, **104**(51), 20404–20409.

VAN DER GUCHT, K., VANDEKERCKHOVE, T., VLOEMANS, N., COUSIN, S., MUYLAERT, K., SABBE, K., GILLIS, M., DECLERK, S., MEESTER, L. & VYVERMAN, W., 2005. Characterization of bacterial communities in four freshwater lakes differing in nutrient load and food web structure. *FEMS Microbiology Ecology*, **53**(2), 205–220.

VAN DER WIELEN, P., BOLHUIS, H., BORIN, S., DAFFONCHIO, D., CORSELLI, C., GIULIANO, L., D'AURIA, G., DE LANGE, G., HUEBNER, A., VARNAVAS, S., THOMSON, J., TAMBURINI, C., MARTY, D., MCGENITY, T., TIMMIS, K. & BIODEEP SCI PARTY, 2005. The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science*, **307**(5706), 121–123.

VAN NIMWEGEN, E., 2003. Scaling laws in the functional content of genomes. *Trends in Genetics*, **19**, 479–484.

VELLEND, M., 2010. Conceptual synthesis in community ecology. *Quarterly Review of Biology*, **85**, 183–206.

VENTER, J., REMINGTON, K., HEIDELBERG, J., HALPERN, A., RUSCH, D., EISEN, J., WU, D., PAULSEN, I., NELSON, K., NELSON, W. ET AL., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**(5667), 66–74.

VON MERING, C., HUGENHOLTZ, P., RAES, J., TRINGE, S. G., DOERKS, T., JENSEN, L. J., WARD, N. & BORK, P., 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**(5815), 1126–1130.

WALTHER, G., POST, E., CONVEY, P., MENZEL, A., PARMESAN, C., BEEBEE, T., FROMENTIN, J., HOEGH-GULDBERG, O., BAIRLEIN, F. ET AL., 2002. Ecological responses to recent climate change. *Nature*, **416**(6879), 389–395.

WANG, Q., GARRITY, G., TIEDJE, J. & COLE, J., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**(16), 5261–5267.

WARNECKE, F., SOMMARUGA, R., SEKAR, R., HOFER, J. & PERNTHALER, J., 2005. Abundances, identity, and growth state of Actinobacteria in mountain lakes of different UV transparency. *Applied and Environmental Microbiology*, **71**(9), 5551–5559.

WEBB, C., 2000. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, **156**(2), 145–155.

WEBB, C., ACKERLY, D., MCPEEK, M. & DONOGHUE, M., 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505.

WENNEKES, P., ROSINDELL, J. & ETIENNE, R., In press. The neutral-niche debate: a philosophical perspective. *Acta Biotheoretica*, doi:10.1007/s10441–012–9144–6.

WHITAKER, R. J., 2006. Allopatric origins of microbial species. *Philosophical Transactions of the Royal Society B - Biological Sciences*, **361**(1475), 1975–1984.

WHITAKER, R. J., GROGAN, D. W. & TAYLOR, J. W., 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, **301**(5635), 976–978.

WHITMAN, W., COLEMAN, D. & WIEBE, W., 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences USA*, **95**(12), 6578–6583.

WHITTAKER, R. H., 1967. Gradient analysis of vegetation. *Biological Reviews*, **42**(2), 207–264.

WIENS, J. & DONOGHUE, M., 2004. Historical biogeography, ecology and species richness. *Trends in Ecology & Evolution*, **19**(12), 639–644.

WILLIS, J. C., 1922. *Age and area: a study in geographical distribution and origin of species*. Cambridge University Press, Cambridge, UK.

WILLNER, D., THURBER, R. & ROHWER, F., 2009. Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology*, **11**(7), 1752–1766.

WOESE, C. R., 1987. Bacterial evolution. *Microbiology and Molecular Biology Reviews*, **51**(2), 221–271.

WOMACK, A., BOHANNAN, B. & GREEN, J., 2010. Biodiversity and biogeography of the atmosphere. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**(1558), 3645–3653.

WRIGHT, I., REICH, P., WESTOBY, M., ACKERLY, D., BARUCH, Z., BONGERS, F., CAVENDER-BARES, J., CHAPIN, T., CORNELISSEN, J., DIEMER, M. ET AL., 2004. The worldwide leaf economics spectrum. *Nature*, **428**, 821–827.

WU, M. & EISEN, J., 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, **9**(10), R151.

WU, X., XI, W., YE, W. & YANG, H., 2007. Bacterial community composition of a shallow hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences. *FEMS Microbiology Ecology*, **61**(1), 85–96.

XU, L., CHEN, H., HU, X., ZHANG, R., ZHANG, Z. & LUO, Z., 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular Biology and Evolution*, **23**(6), 1107–1108.

YOUSSEF, N. & ELSHAHED, M., 2008. Diversity rankings among bacterial lineages in soil. *The ISME Journal*, **3**, 305–313.

ZABALLOS, M., LÓPEZ-LÓPEZ, A., OVREAS, L., BARTUAL, S., D'AURIA, G., ALBA, J., LEGAULT, B., PUSHKER, R., DAAE, F. & RODRÍGUEZ-VALERA, F., 2006. Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. *FEMS Microbiology Ecology*, **56**(3), 389–405.

ZANEVELD, J., LOZUPONE, C., GORDON, J. & KNIGHT, R., 2010. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research*, **38**(12), 3869–3879.